

Course #412

Analyzing Microarray Data using the mAdb System

February 16-17, 2005 1:00 pm - 4:00pm

madb-support@bimas.cit.nih.gov

- Intended for users of the mAdb system who are familiar with mAdb basics
- Focus on analysis of multiple array experiments

Esther Asaki, Liming Yang, John Powell

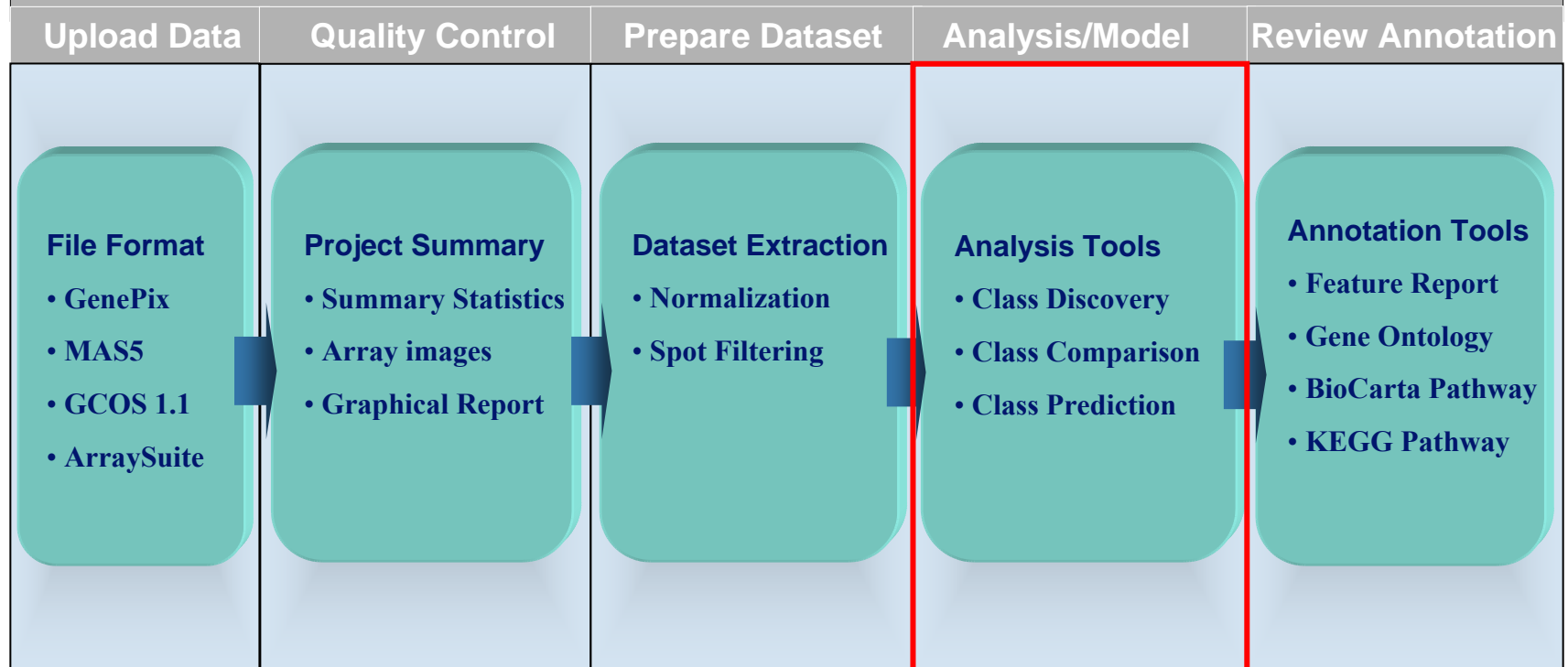
Agenda

1. mAdb system overview
2. mAdb dataset overview
3. mAdb analysis tools for dataset
 - Class Discovery - clustering, PCA, MDS
 - Class Comparison - statistical analysis
 - t-test
 - ANOVA
 - Significance Analysis of Microarrays - SAM
 - Class Prediction - PAM

Various Hands-on exercises

1. mAdb system overview

mAdb Data Workflow




2. mAdb dataset overview

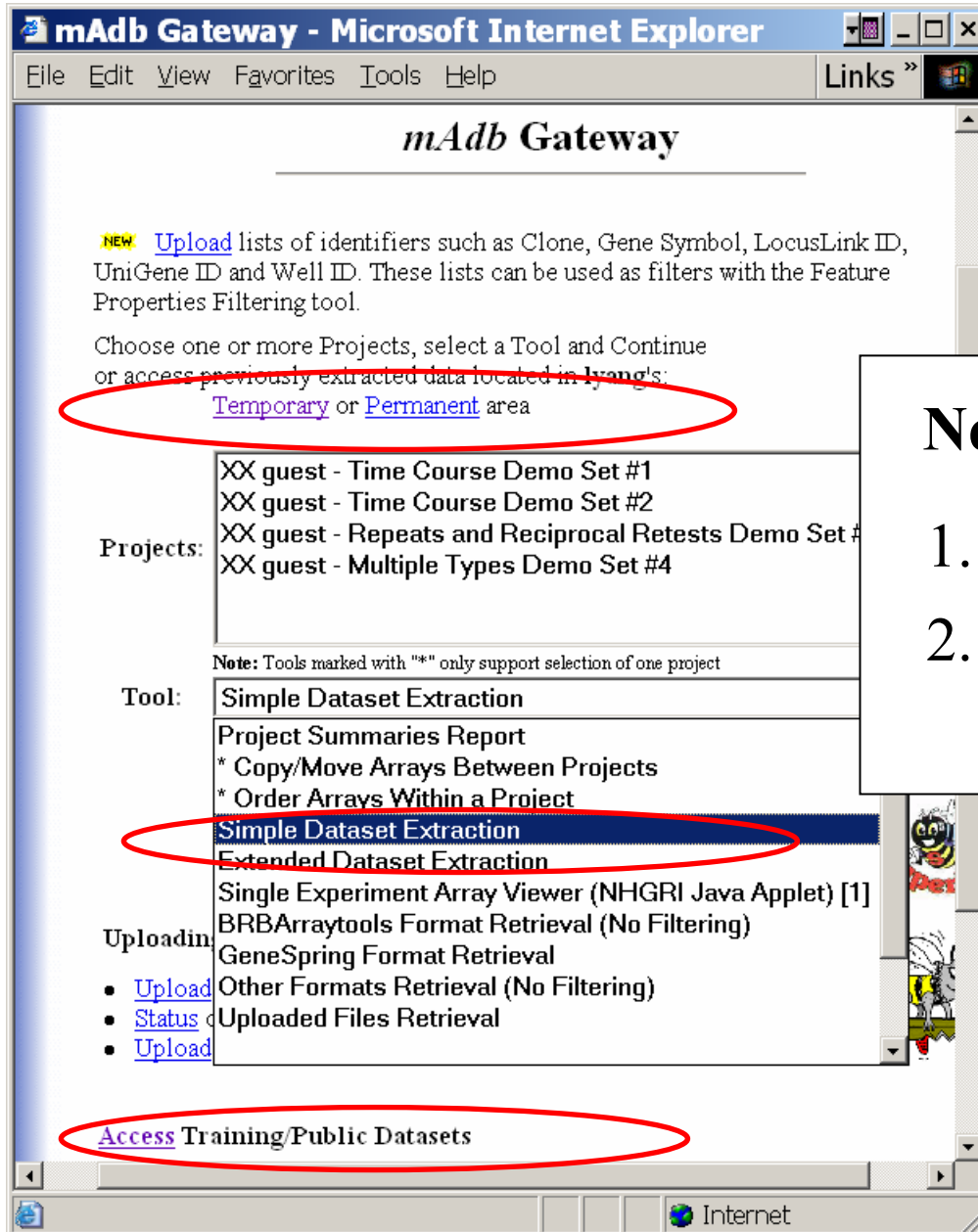
What is a dataset?

- mAdb Dataset
 - Collection of data from multiple experiments
 - Genes as rows and experiments as columns

Genes		sample1	sample2	sample3	sample4	sample5	...
	1	0.46	0.30	0.80	1.51	0.90	...
	2	-0.10	0.49	0.24	0.06	0.46	...
	3	0.15	0.74	0.04	0.10	0.20	...
	4	-0.45	-1.03	-0.79	-0.56	-0.32	...
	5	-0.06	1.06	1.35	1.09	-1.09	...



Gene expression level = (normalized) Log(Red signal / Green signal)



New or Existing Dataset:

1. Create New Dataset
2. Access Existing Dataset

A	0.008	1.	HDLM2_A	HL_HDLM2
A	0.007	2.	JIM3_A	MM_JIM3
A	0.007	3.	JJN3_A	MM_JJN3
A	0.006	4.	L428_A	HL_L428
A	0.009	5.	L540_A	HL_L540
A	0.006	6.	Ly10_A	DLBCL_Ly10
A	0.007	7.	Ly19_A	DLBCL_Ly19
A	0.007	8.	Ly3_A	DLBCL_Ly3
A	0.007	9.	Ly7_A	DLBCL_Ly7
A	0.007	10.	U266_A	MM_U266

[Edit](#) Data for Dataset: **Cell Lines representing 3 Lymphomas**

10 Arrays and 22283 Expression Rows extracted.
 Data transformation method: Centered to Signal Median
 Spot Filter Options:
 Signals are floored at 100.0

[Expand](#) this Dataset.
 Access Datasets in your [Temporary](#) area.

Dataset Display Page

- Dataset History
- Analysis Tools
- Retrieval and Display Options...

[Filtering/Grouping/Analysis Tools](#)

Choose a Tool Additional Filtering Options and Proceed

[Interactive Graphical Viewers](#)

Choose a Viewer MDS: MultiDimensional Scaling and View

[Dataset Retrieval & Display Options](#)

Retrieve Dataset formatted for Eisen Cluster

Redisplay ☒ Show Array Details at the top of the page

Dataset Display

Redisplay

☒ Show Array Details at the top of the page

Background Color - None - Contrast 1.585

Limiting display to to 25 genes

☒ Show Data Values ☒ Use Names in Column Heading

☐ Apply log2 transform ☐ Use Description in Column Heading

☒ Show Gene Symbols ☐ Show Map Information

☐ Show UniGene Cluster ☐ Show BioCarta Pathways

☐ Show KEGG Pathways

☐ Show GO Tier 2 Component ☐ Show GO Tier 3 Component

☐ Show GO Tier 2 Function ☐ Show GO Tier 3 Function

☐ Show GO Tier 2 Process ☐ Show GO Tier 3 Process

☒ Show Gene Description ☐ Show GO Terms

☐ Show Average(Log2 Ratio) ☐ Show Max(Log2 Ratio)-Min(Log2 Ratio)

☐ Show Variance

[Save](#) a Feature Property List (used with the Feature Properties Filtering tool).

Records 1 to 25 of 22283 total records displayed.

A	A	A	A	A	A	A	A	A	A	⬇	⬆	⬇	⬆	⬇	⬆
HDLM2_A	JIM3_A	JJN3_A	L428_A	L540_A	Ly10_A	Ly19_A	Ly3_A	Ly7_A	U266_A	Well ID	Feature ID	Gene			
0.8986	1.1075	0.8887	1.5182	1.1664	1.3198	1.2333	0.6761	0.8685	0.9967	1118566	117_at	HSPA6			
8.1537	6.7782	8.5125	6.8697	9.1886	7.6118	9.1357	7.4983	8.7316	5.8007	1118567	121_at	PAX8			

- Dataset display options dynamic
- Integrated gene information
- Newly created dataset puts all experiments into a single group

mAdb Dataset Display

Group label
Sample name

genes

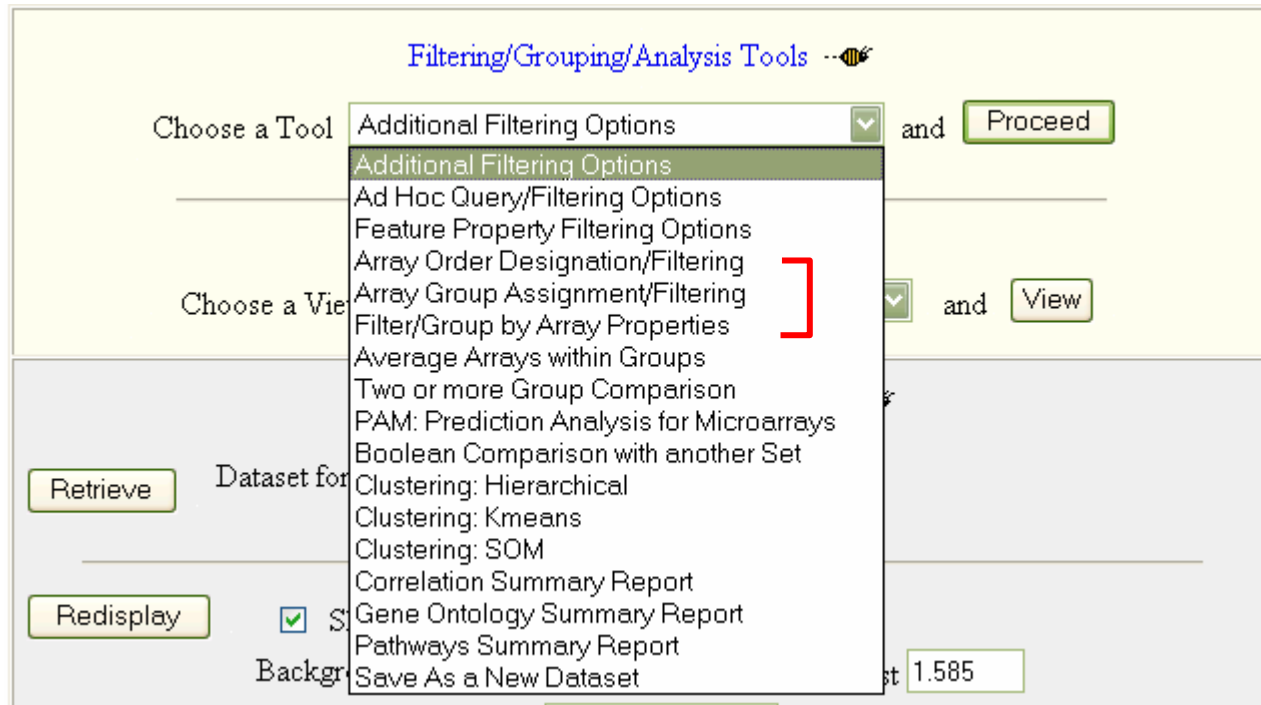
A	A	A	A	A
BJAB_A_B	Daudi_A_B	Jurkat_A_B	Ly10_A_B	Ly3_A_B
			7.7702	
9.7305	9.7985	9.7249	10.2981	10.1150
	8.9715			
	8.8918	9.0752	10.2200	
8.4250	7.0224	7.8511	7.4692	7.7886
6.9189	7.5645			7.7814
9.3296	9.6202	9.4409	9.9652	10.0534
			7.8629	7.3505
10.0053	9.6605	9.3872	9.9003	9.3181
8.1908	8.2187	7.3540	8.3650	
6.5014			7.0629	
	6.5251	6.4512		
9.6604	10.0402	8.6991	9.9747	9.4539
8.3781	8.8981	8.1739	8.2322	9.3807
7.9419	7.4741	7.9301		
8.9372	9.8243	9.4774	9.7465	10.2738
8.2002			9.9105	9.6255
5.0575	6.8163	5.9542		5.7388
9.9564	9.8420	9.7677	10.1529	9.3419
9.9284	9.6363	9.3726	9.8858	10.1808
9.4419	9.0507	9.4075	9.9434	9.0739
10.4035	9.7502	9.2389	10.1029	10.5434
9.0906	9.3452	9.3869	9.6770	9.3613

Well ID	Feature ID	Gene	Description
1118566	117_at	HSPA6	heat shock 70kDa protein 6 (HSP70B')
1118567	121_at	PAX8	paired box gene 8
1118568	177_at	PLD1	phospholipase D1, phosphatidylcholine-sp
1118569	179_at	PMS2L9	postmeiotic segregation increased 2-like
1118570	320_at	PEX6	peroxisomal biogenesis factor 6
1118572	564_at	GNA11	guanine nucleotide binding protein (G pr
1118573	632_at	GSK3A	glycogen synthase kinase 3 alpha
1118574	823_at	CX3CL1	chemokine (C-X3-C motif) ligand 1
1118575	1053_at	RFC2	replication factor C (activator 1) 2, 40kD
1118576	1294_at	UBE1L	ubiquitin-activating enzyme E1-like
1118577	1316_at	THRA	thyroid hormone receptor, alpha (erythro
1118579	1431_at	CYP2E1	cytochrome P450, family 2, subfamily E
1118581	1487_at	ESRRA	estrogen-related receptor alpha
1118582	1729_at	TRADD	TNFRSF1A-associated via death domain
1118584	1861_at	BAD	BCL2-antagonist of cell death
1118585	243_g_at	MAP4	microtubule-associated protein 4
1118586	266_s_at	CD24	CD24 antigen (small cell lung carcinoma
1118587	31799_at		Sapiens clone 24627 mRNA sequence
1118588	31807_at	DDX49	DEAD (Asp-Glu-Ala-Asp) box polypepti
1118589	31826_at	KIAA0674	KIAA0674 protein
1118591	31837_at	BC002942	hypothetical protein BC002942
1118592	31845_at	ELF4	E74-like factor 4 (ets domain transcripti
1118594	31861_at	IGHMBP2	immunoglobulin mu binding protein 2

Dataset Group Assignment

- Array Order Designation/Filtering
- Array Group Assignment/Filtering
- Filter/Group by Array Properties

Dataset group assignment tools



Array Order Designation/Filtering

Change Array order.

↑
↓

Arrays Included

- HDLM2_A HL_HDLM2
- L428_A HL_L428
- L540_A HL_L540
- JIM3_A MM_JIM3
- JJN3_A MM_JJN3
- U266_A MM_U266
- Ly10_A DLBCL_Ly10
- Ly19_A DLBCL_Ly19
- Ly3_A DLBCL_Ly3
- Ly7_A DLBCL_Ly7

↓ Remove or Add Back Arrays ↑

Arrays Excluded

Subset Label:

- Order arrays in dataset
- Delete/Add back arrays in dataset
- Subsequent analysis will be ordered by groups first and then ordered within each group
- Does not group arrays

Array Group Assignment/Filtering

Note the --🐛 marks items which lead to additional help when clicked

Dataset Properties --🐛

Subset Label:

Expand the number of possible Group Designations to 4, 5, 6, 7, 8, 16 or 24 groups.

Group Designation --🐛

--	A	B	C	Submit	Cancel
	A	B	C	Array Name & Description	
<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	HDLM2_A HL_HDLM2	
<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	JIM3_A MM_JIM3	
<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	JJN3_A MM_JJN3	
<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	L428_A HL_L428	
<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	L540_A HL_L540	
<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	Ly10_A DLBCL_Ly10	
<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	Ly19_A DLBCL_Ly19	
<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	Ly3_A DLBCL_Ly3	
<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	Ly7_A DLBCL_Ly7	
<input type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	U266_A MM_U266	

- One click per array for additional group
- Not convenient for large dataset
- Can not order within group

Filter/Group by Array Properties

mAdb Dataset Display

A	0.008	1.	HDLM2_A	HL_HDLM2
A	0.007	2.	JIM3_A	MM_JIM3
A	0.007	3.	JJN3_A	MM_JJN3
A	0.006	4.	L428_A	HL_L428
A	0.009	5.	L540_A	HL_L540
A	0.006	6.	Ly10_A	DLBCL_Ly10
A	0.007	7.	Ly19_A	DLBCL_Ly19
A	0.007	8.	Ly3_A	DLBCL_Ly3
A	0.007	9.	Ly7_A	DLBCL_Ly7
A	0.007	10.	U266_A	MM_U266

[Edit](#) Data for Dataset: Cell Lines representing 3 Lymphomas

10 Arrays and 22283 Expression Rows extracted.
Data transformation method: Centered to Signal Median
Spot Filter Options:
Signals are floored at 100.0

- Array properties include Name and Short Description
- Identify consistent pattern

Filter/Group by Array Properties

The screenshot shows a web interface for filtering and grouping arrays. It features five groups (A through E) with dropdown menus for selecting properties and comparison operators. Group A and B use 'Short Description' and 'Begins with' with values 'HL' and 'MM' respectively. Group C uses 'Short Description' and 'Begins with' with value 'DLBCL'. Group D and E use 'Array Name'. A dropdown menu is open for Group D, showing options: 'Contains', 'Begins with' (highlighted), 'Equals', 'Does Not Contain', 'Does Not Begin with', and 'Does Not Equal'. Below the groups, a text input field for 'Subset Label' contains 'Filter/Group by Array Property'. At the bottom are 'Submit' and 'Cancel' buttons.

Group	Property	Operator	Value
Group A	Short Description	Begins with	HL
Group B	Short Description	Begins with	MM
Group C	Short Description	Begins with	DLBCL
Group D	Array Name	Begins with	
Group E	Array Name		

Expand the number of possible Group Designations to 10 , 15 , 20 or 26 groups.

Subset Label: Filter/Group by Array Property

Submit Cancel

- Convenient for large dataset
- Can not order arrays within group

Group Assignment

A	A	A	B	B	B	C	C	C	C	↓	↑	↓	↑	↓	↑
HDLM2_A	L428_A	L540_A	JIM3_A	JJN3_A	U266_A	Ly3_A	Ly7_A	Ly10_A	Ly19_A	Well ID	Feature ID	Gene			
0.8986	1.5182	1.1664	1.1075	0.8887	0.9967	0.6761	0.8685	1.3198	1.2333	1118566	117_at	HSPA6			
8.1537	6.8697	9.1886	6.7782	8.5125	5.8007	7.4983	8.7316	7.6118	9.1357	1118567	121_at	PAX8			
0.8042	2.2147	0.8831	0.6680	0.6954	1.4118	0.6761	0.6743	0.6046	0.7337	1118568	177_at	PLD1			
4.1856	6.4728	9.8080	5.3601	6.0779	5.1954	7.1981	3.7505	7.2110	4.8481	1118569	179_at	PMS2L9			
2.3557	1.6427	1.2628	2.5865	2.4068	2.0954	1.4949	2.1160	1.0713	2.5561	1118570	320_at	PEX6			
1.1856	1.3852	0.9514	0.9599	0.9757	0.8588	1.2529	1.4626	1.3452	1.2318	1118571	336_at	TBXA2R			
3.7746	1.6271	2.5043	1.1516	1.0508	0.6536	1.4875	1.9670	1.1227	1.1988	1118572	564_at	GNA11			
4.5008	5.1783	5.5333	5.3079	7.4172	6.8863	7.1846	5.8658	6.0435	8.4519	1118573	632_at	GSK3A			
4.1646	12.1329	0.8532	0.6680	0.6954	0.6536	1.1034	0.6743	1.4075	0.7337	1118574	823_at	CX3CL1			
5.5663	4.3223	5.4480	1.6206	2.9270	4.4418	4.3158	3.3790	5.7775	3.3067	1118575	1053_at	RFC2			
3.9173	2.4157	2.0461	1.3460	0.9437	1.1039	1.3083	2.0964	1.9933	1.9391	1118576	1294_at	UBE1L			
0.7800	0.7918	0.8532	0.7715	0.6954	0.8327	0.6761	0.8483	0.8083	0.7630	1118577	1316_at	THRA			
0.7800	0.6485	0.8532	0.6680	0.6954	0.6536	0.6761	0.6743	0.6046	0.7337	1118578	1320_at	PTPN21			

- Group assignment information is carried into relevant analysis
- Dataset is independent from microarray platforms

Examples for using group labels

- Additional Filtering per Group
- Correlation Summary Report
- Average Arrays within Groups

Filter by Group Properties

Missing Value Filters ..🐝

☒ Genes: Require values in \geq 80 % of Arrays

☐ Arrays: Require values in \geq 30 % of Genes per Group

Gene Filters ..🐝

☐ Ratio \geq 2 in \geq 80 % of Arrays
 ☒ *Apply Symmetrically*

☐ Ratio \geq 2 in \geq 50 % of Arrays OR
 Ratio \leq 0.5 in \geq 50 % of Arrays

☐ Average Ratio \geq 0
 ☐ *Apply Symmetrically*



















☐ Max (Ratio) / Min (Ratio) \geq 1.2

☐ Variance (Gene Vector) percentile \geq 90 %

- Ensures each group has sufficient number of non-missing values

Correlation Summary Report

Correlations

A	A	A	B	B	B	C	C	C	C							
#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Grp			Array Name	Array Description		
#1 A	0.890	0.914	0.844	0.873	0.852	0.853	0.838	0.856	0.836	A			1.	HDLM2_A	HL_HDLM2	
	#2 A	0.882	0.852	0.860	0.847	0.856	0.824	0.869	0.845	A			2.	L428_A	HL_L428	
		#3 A	0.860	0.880	0.855	0.858	0.850	0.859	0.843	A			3.	L540_A	HL_L540	
			#4 B	0.896	0.895	0.852	0.826	0.850	0.846	B			4.	JIM3_A	MM_JIM3	
				#5 B	0.885	0.868	0.853	0.859	0.867	B			5.	JJN3_A	MM_JJN3	
					#6 B	0.857	0.832	0.852	0.848	B			6.	U266_A	MM_U266	
						#7 C	0.871	0.924	0.882	C			7.	Ly10_A	DLBCL_Ly10	
							#8 C	0.873	0.918	C			8.	Ly19_A	DLBCL_Ly19	
								#9 C	0.883	C			9.	Ly3_A	DLBCL_Ly3	
									#10 C	C			10.	Ly7_A	DLBCL_Ly7	

- Pair wise correlation between 2 samples in dataset
- Individual scatter plot available
- Group pattern for quality control

[Home Page](#) | [mAdb Gateway](#) | [Upload Status](#)
[Forums](#) | [Reference Info](#) | [Program Downloads](#) | [GeneCards](#)

mAdb Correlation Report

View Array Summaries

[Return](#) to the input dataset.

Redisplay

Background Color Scheme **Green/White/Red**

Color Saturation Max/Mid/Min **0.8** **0.6** **0.4**

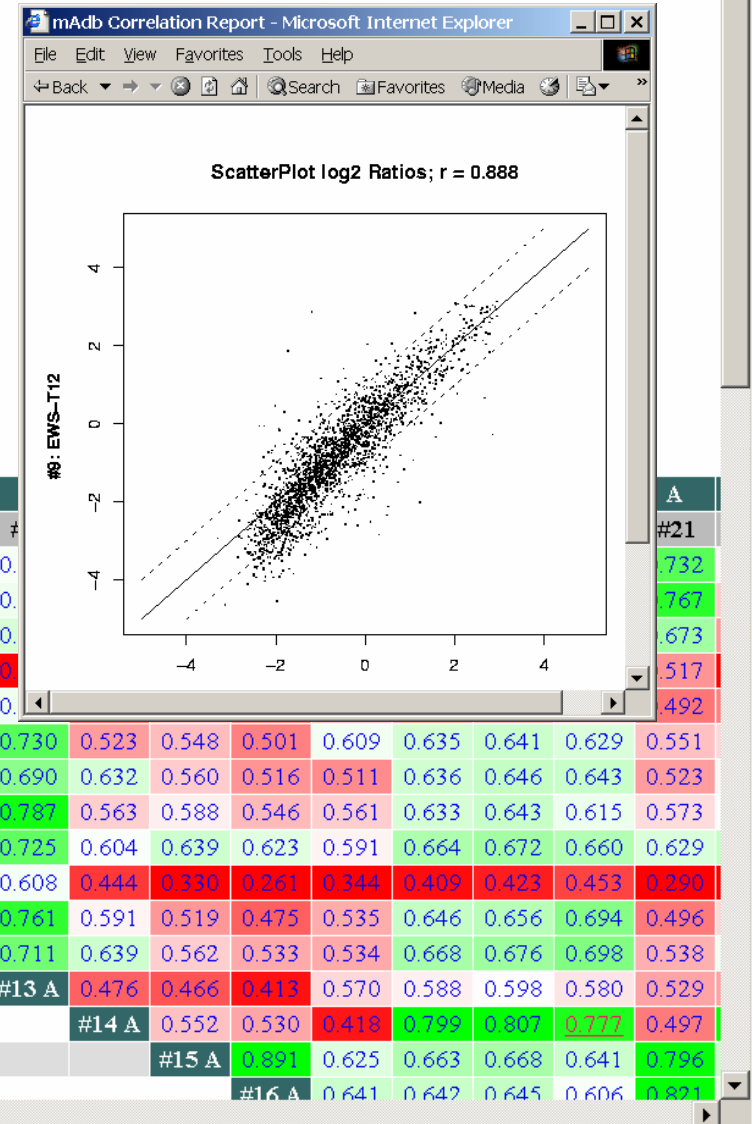
Note: For proper coloring Max > Mid > Min

Note: Click on the Correlation values to display the corresponding ScatterPlot


Correlations

	A	A	A	A	A	A	A	A	A	A	A	A																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																			
--	---	---	---	---	---	---	---	---	---	---	---	---	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--


Visual Bivariate Data Analysis



Average Arrays within Group

Filtering/Grouping/Analysis Tools 

Choose a Tool and

Interactive Graphical Viewers 

Choose a Viewer and

- Averages using log ratios - though user chooses to display linear or log2 values

Dataset I

Small Round Blue Cell Tumors (SRBCTs)

- Khan et al. *Nature Medicine* 2001
- 4 tumor classifications
- 63 training samples, 25 testing samples, 2308 genes
- Neural network approach

Hands-on Session 1

- Lab 1- Lab 4
- Read the questions before starting, then answer them in the lab.
- Use web site: <http://mAdb-training.cit.nih.gov>
- Avoid maximizing web browser to full screen.
- Total time: 20 minutes

3. mAdb dataset analysis tools

- Class Discovery: clustering, PCA, MDS
- Class Comparison: statistical analysis
- Class Prediction: PAM

Analysis Overview

Class Discovery - Unsupervised	<ul style="list-style-type: none"> • Clustering – Hierarchical, K-means, SOMs • Principal components Analysis (PCA) • Multidimensional Scaling (MDS)
Class Comparison - Supervised	<ul style="list-style-type: none"> • paired t-tests • t-test pooled (equal) variance • t-test separate (unequal) variance • Significance Analysis of Micro- arrays (SAM) • One way ANOVA • Wilcoxon Rank-Sum (Mann Whitney U) • Wilcoxon Matched-pairs Signed Rank • Kruskal-Wallis
Class Prediction - Supervised	Prediction Analysis for Microarrays (PAM)

Class Discovery Example

- Discover cancer subtypes by gene expression profiles
- Identify genes which have different expression patterns in different groups
- Tools: Cluster Analysis, PCA and MDS

Class Comparisons Example

- Find genes that are differentially expressed among cancer groups
- Find genes up/down regulated by drug treatment
- Tools:
 - Two or more group comparison
 - Statistics Results filtering

Class Prediction Example

- Identify an expression profile which correlates with survival in certain cancers
- Identify an expression profile which can be used to diagnose different types of lymphomas
- Tools: Prediction Analysis for Microarrays (PAM)

3. mAdb dataset analysis tools

- Class Discovery: clustering, PCA, MDS
- Class Comparison: statistical analysis
- Class Prediction: PAM

Class Discovery

- Dataset with large amount of data
- Dataset not organized
- Visualization with Clustering, PCA, MDS

Cluster Analysis

- Organize large microarray dataset into meaningful structures
- Visualize and extract expression patterns

What to Cluster?

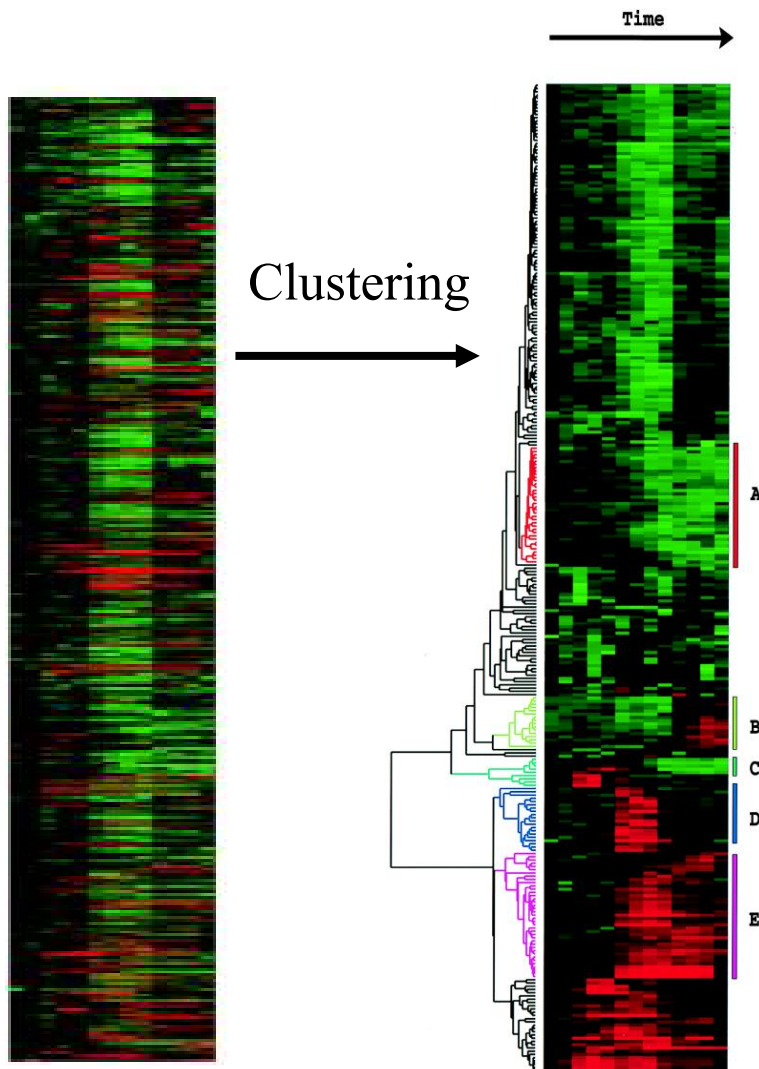
Genes - identify groups of genes that have correlated expression profiles

Samples - put samples into groups with similar overall gene expression profiles

Clustering Methods

- Hierarchical clustering
- Partitional clustering
 - K-means
 - Self-Organizing Maps (SOM)

Cluster Example on Genes



Much easier to look at large blocks of similarly expressed genes

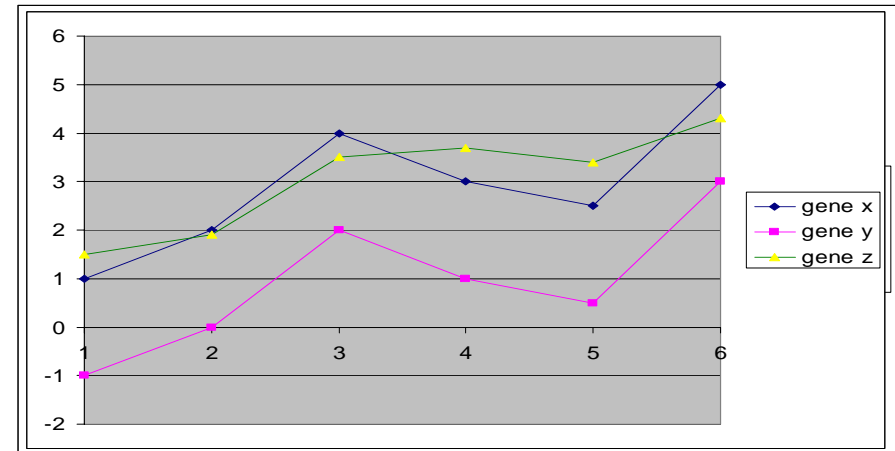
Dendrogram helps show how 'closely related' expression patterns are

- A. Cholesterol syn.
- B. Cell cycle
- C. Immediate-early response
- D. Signaling
- E. Tissue remodeling

2 Steps

– Pick a distance method

- Correlation
- Euclidian

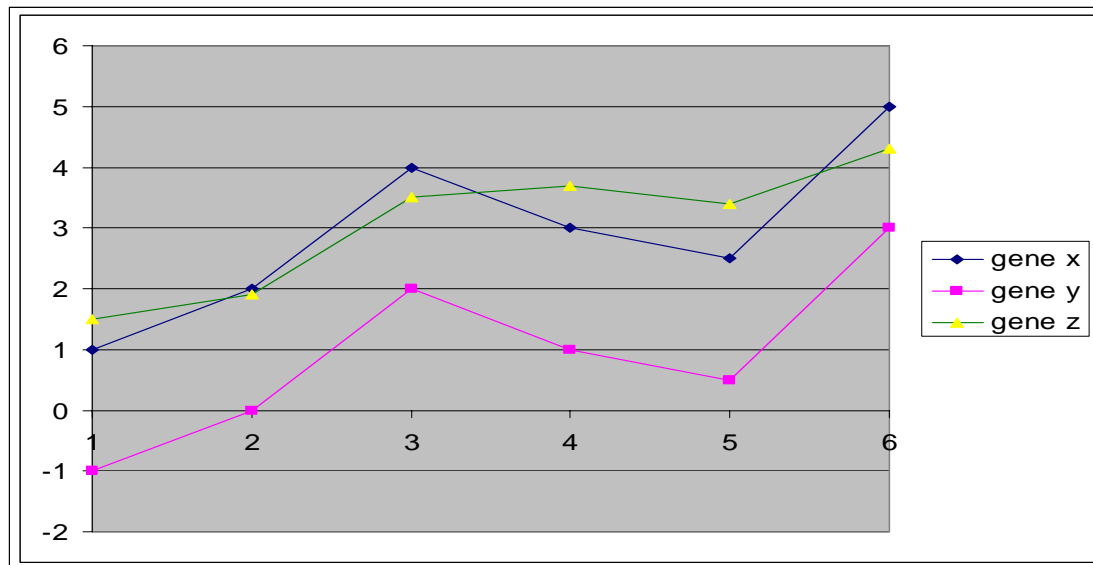


– Pick the linkage method

- Average linkage
- Complete linkage
- Single linkage

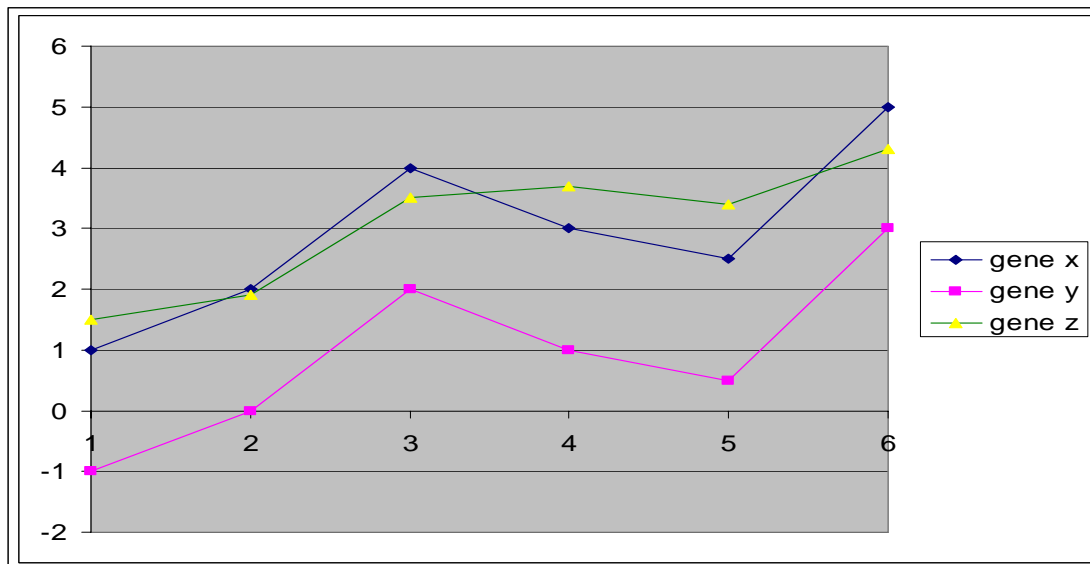
Correlation

- Compares shape of expression curves (-1 to 1)
- Can detect inverse relationships (absolute correlation)

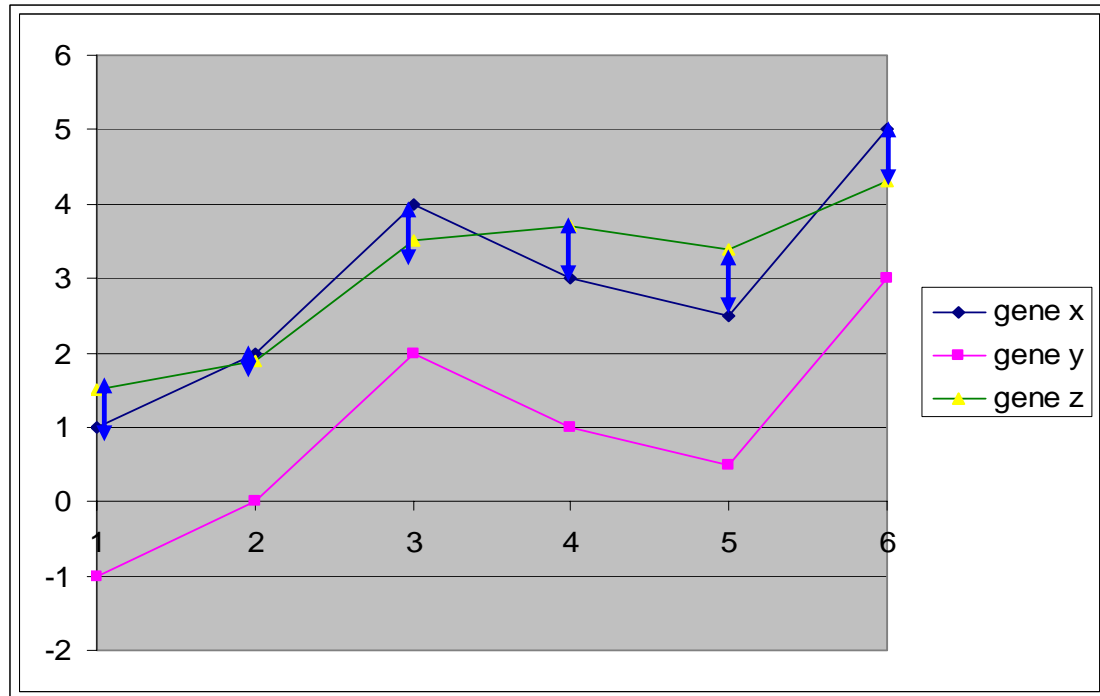


Two Flavors of correlation


- Correlation (centered-classical Pearson)
- Correlation (un-centered)
 - assume the mean of the data is 0, penalize if not
 - Measures both similarity of shape and the offset from 0



Euclidean Distance



Similarity/Distance Metric Summary

Hierarchical Clustering Options 

Similarity/Distance Metric

Genes:

Arrays:

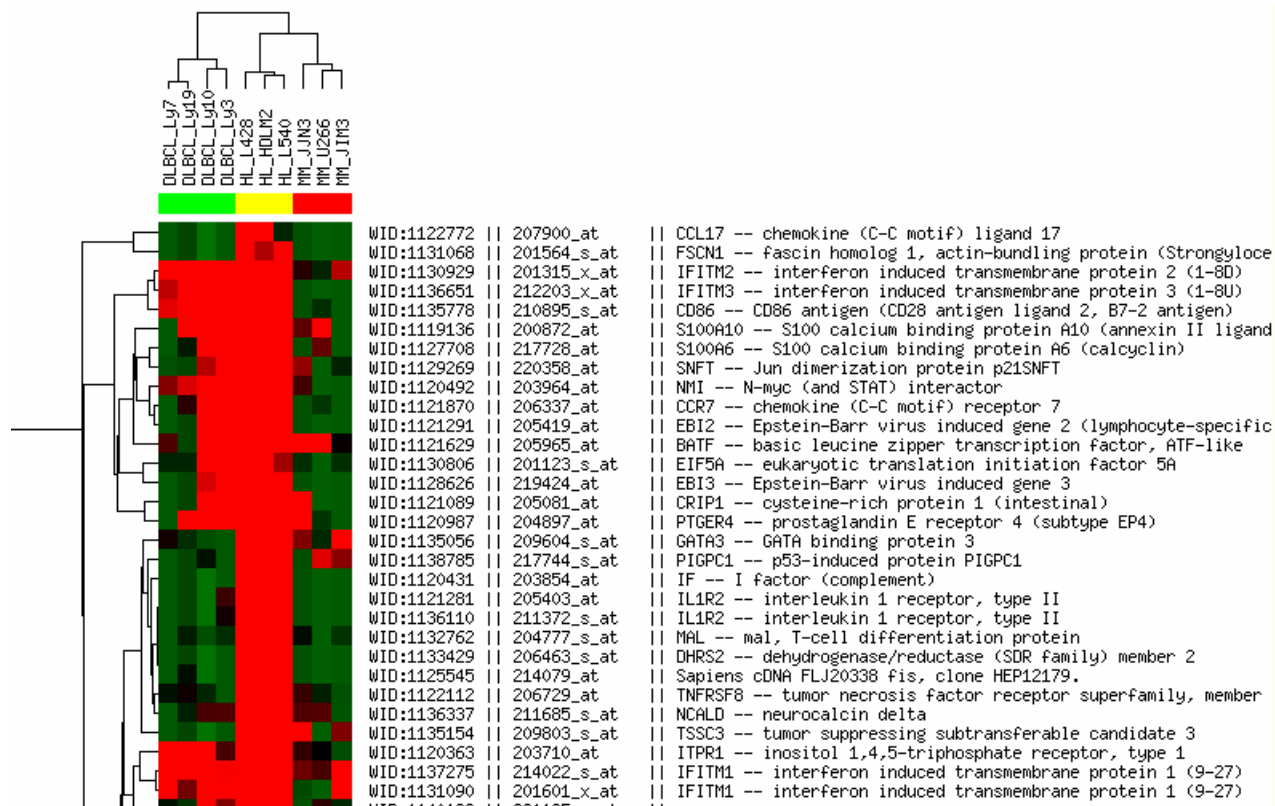
Linkage Method:

shape

Shape and offset

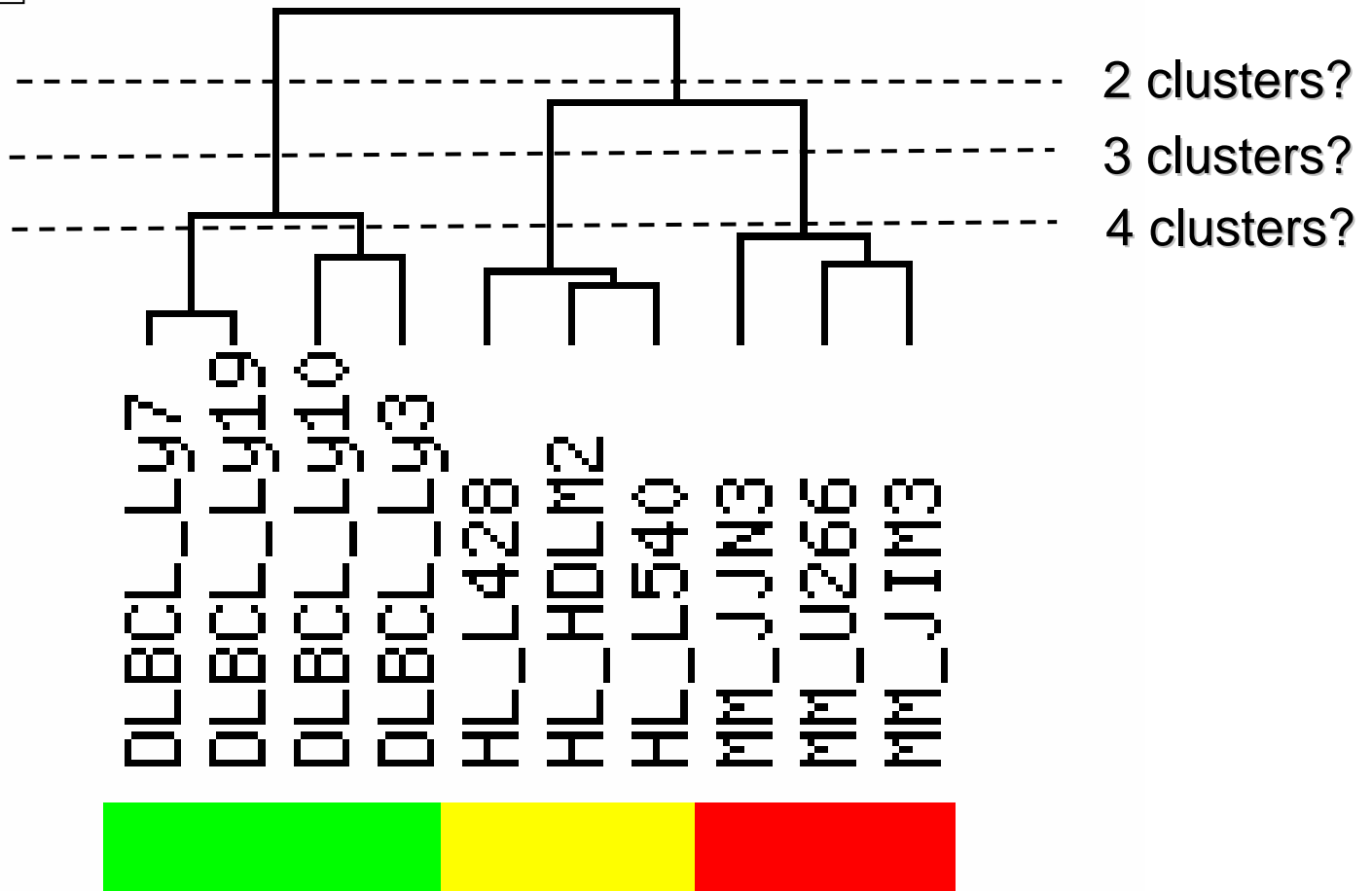
distance

Hierarchical Clustering Example



Degrees of
dissimilarity

Tree Cutting



Hierarchical Clustering Summary

- Detection of patterns for both genes and samples
- Good visualization with tree graphs
- Dataset size limitations
- No partition in results, require tree cutting

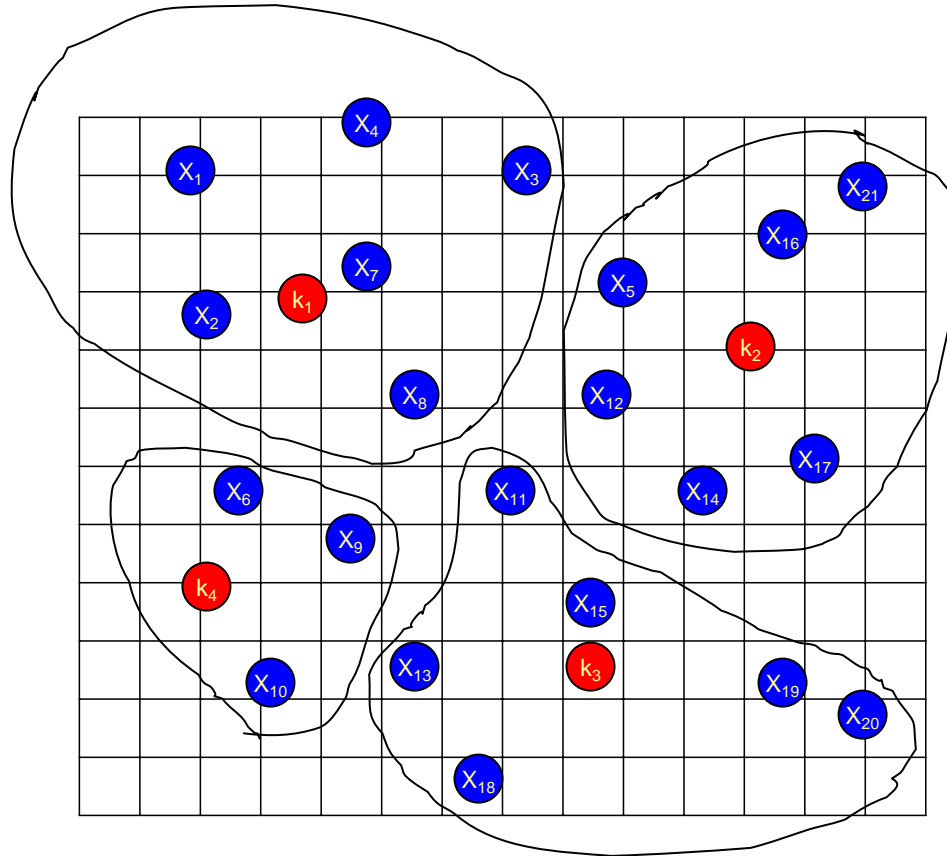
Partitional clustering : K-means

- Partition data into K clusters, with number K supplied by user.
- Produce cluster membership as results.

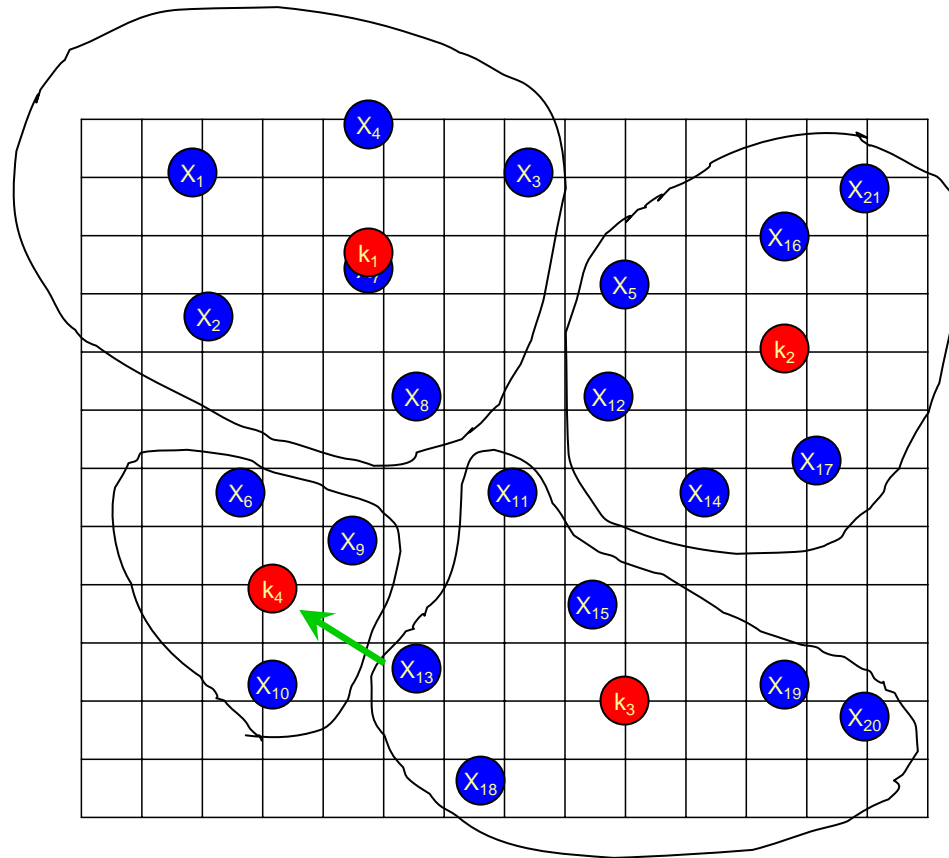
K-means Algorithm

- Divide observations into K clusters.
- Use cluster averages (means) to represent clusters
- Maximize the inter-cluster distance
Minimize intra-cluster distance.

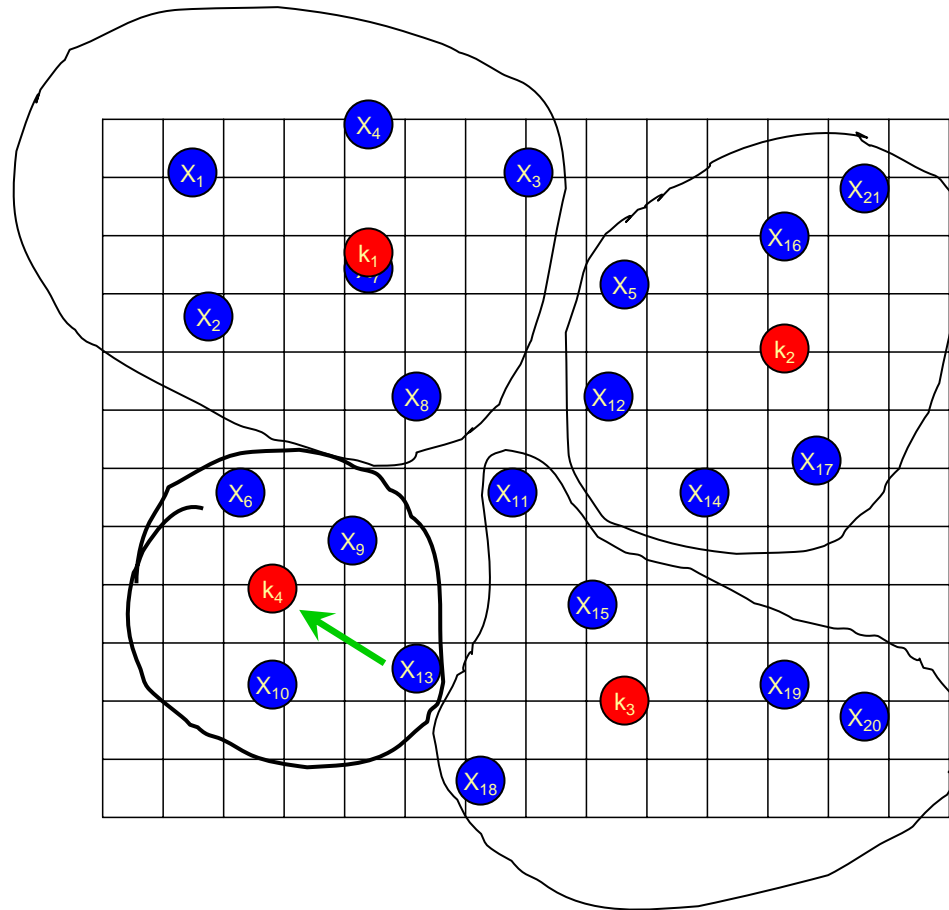
K-means Algorithm



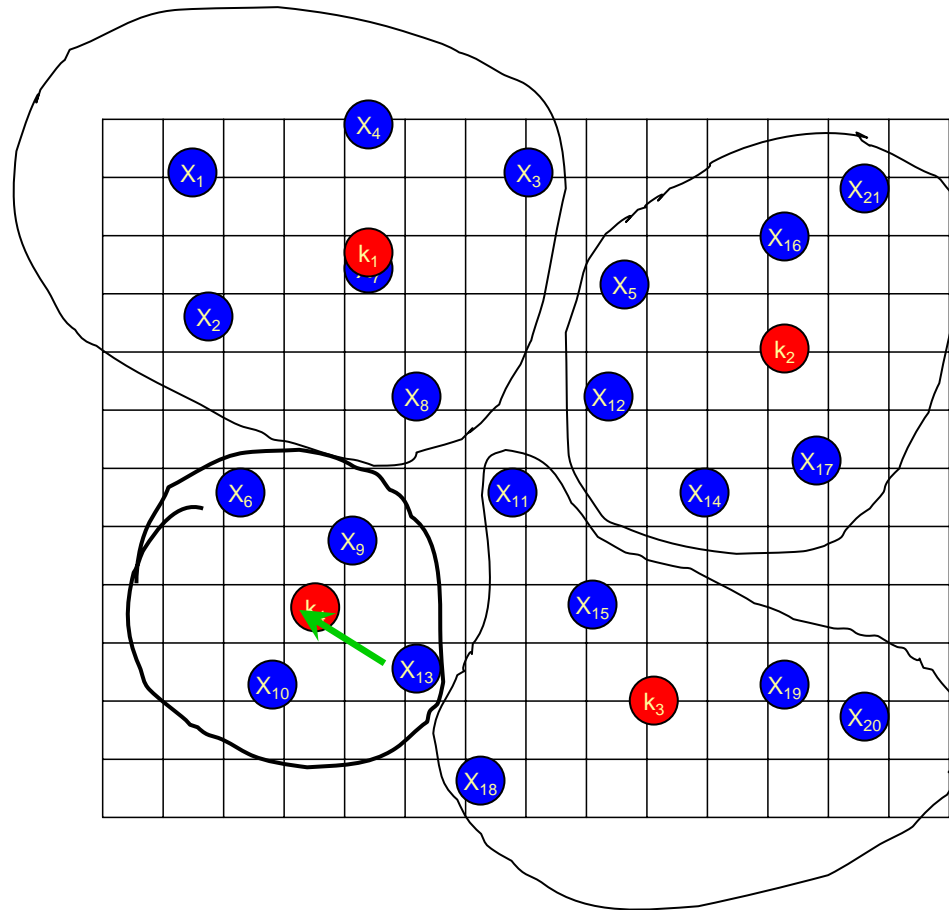
K-means Algorithm



K-means Algorithm



K-means Algorithm




mAdb K-means Options

Set number of clusters




Set number of iteration



Kmeans Clustering Options 

Specify Number of Nodes

Maximum Number of iterations

Kmeans Nodes
Hierarchical Clustering Options 

Similarity/Distance Metric

Genes:

Arrays:

Linkage Method:

Hierarchical clustering
within node

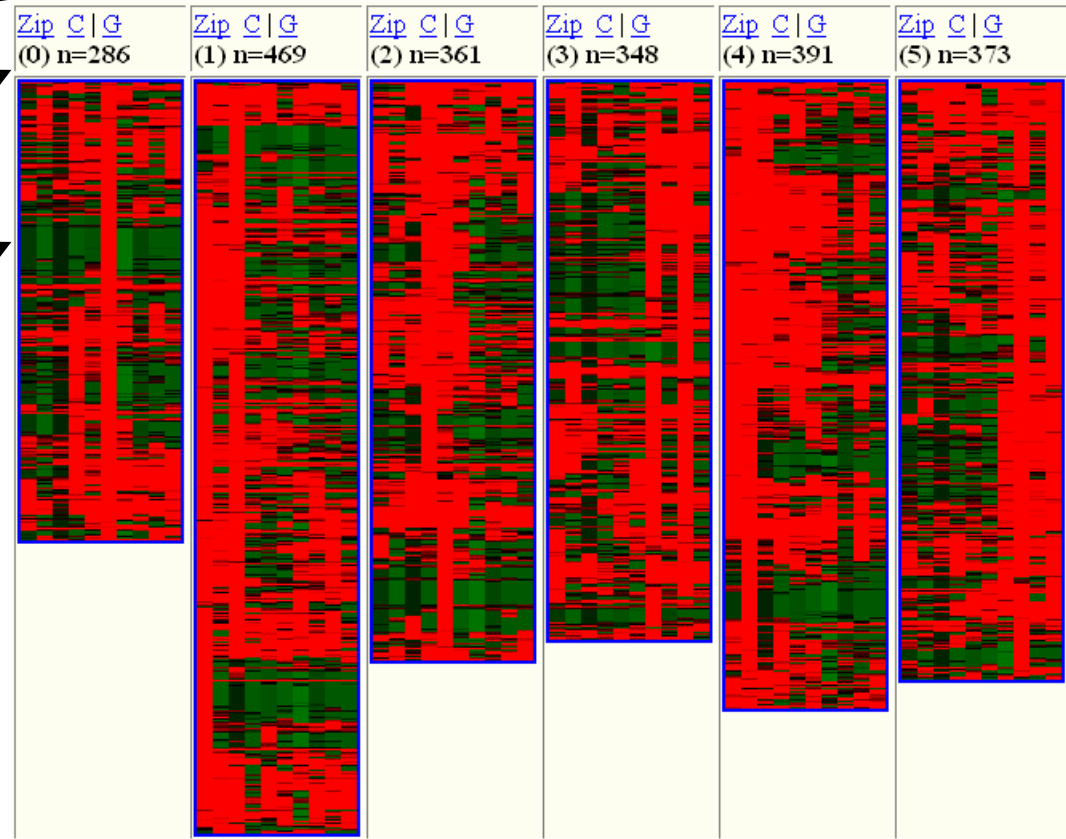


K-means Clustering Example

Save as input to TreeView

Create new subset of genes

Show hierarchical clustering



Summary

- Fast algorithm
- Partitions features into smaller, manageable groups
- mAdb allows hierarchical clustering within each K-mean cluster
- Must supply reasonable number of K
- No relationship among partitions

Self-Organizing Maps (SOM)

- Partitions data into 2 dimensional grid of nodes
- Clusters on the grid have topological relationships
- 2 numbers for the dimension of grid supplied by user

mAdb SOM options

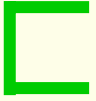
Set number of clusters (X, Y) →

Set number of iteration →

Activate Randomized Partition →

Hierarchical within SOM clusters →

Self Organizing Maps Options



Specify X dimension

Specify Y dimension

Number of iterations

Initialize with Randomized Partition ☒

SOM Elements

Hierarchical Clustering Options

Similarity/Distance Metric

Genes:

Arrays:

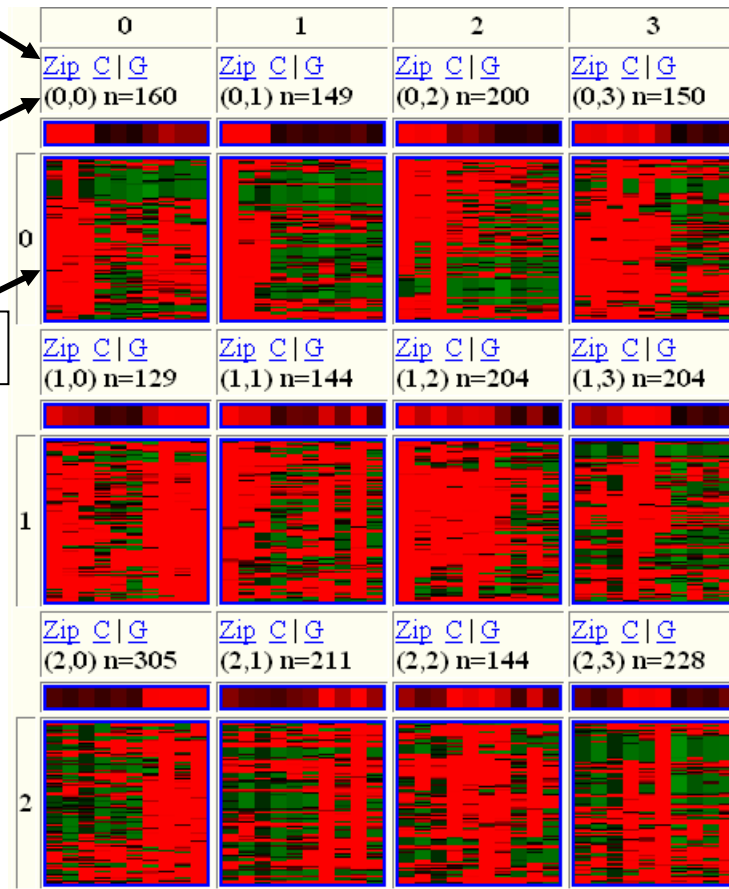
Linkage Method:

SOM Clustering Example

Save as input to TreeView

Create new subset of genes

Show hierarchical clustering



SOM Summary

- Neighboring partitions similar to each other
 - Partitions features into smaller groups
 - mAdb allows hierarchical clustering within each SOM cluster
-
- Results may depend on initial partitions

Summary of mAdb Clustering Tools

	Hierarchical	K-means	SOM
Relationship visualization	Tree Structure	partition Membership	Partition 2-D topology
Data Size	Small	Large	Large
Performance	Slow	Fast	Middle
Cluster Type	Gene/Array	Gene	Gene

Cluster Analysis

- Normalization is important
- Reduce data points by variance
- Use K-mean or SOM to partition dataset
- Use biological information to interpret results

Hands-on Session 2

- Lab 5 - lab 6 (Lab 7 optional)
- Total time: 15 minutes

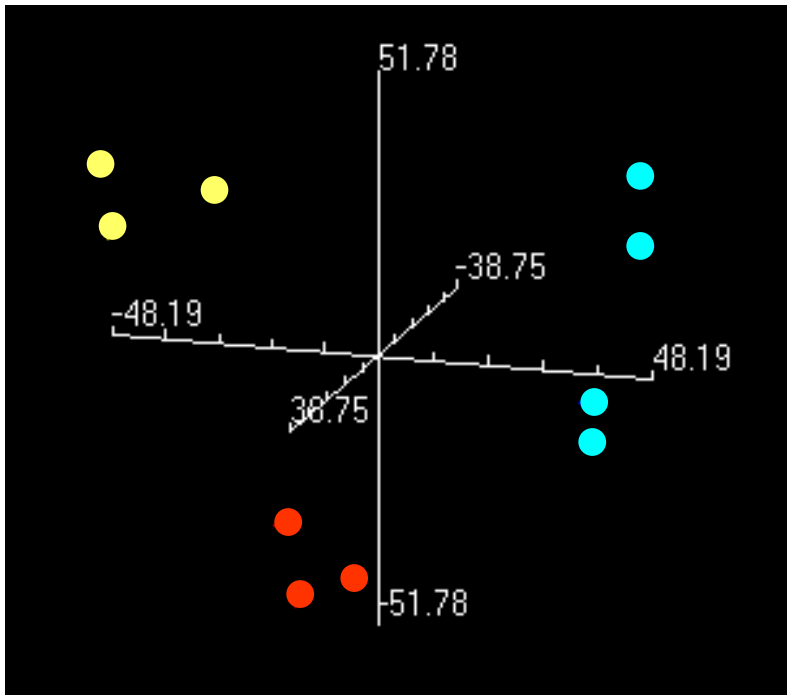
Principal Component Analysis

- How different samples are from each other
- Project high-dimensional data into lower dimensions, which captures most of the variance
- Display data in 2D or 3D plot to reveal the data pattern

Principal Component Analysis

- Hypothesis - there exist unobservable or “*hidden*” variables (complex traits) which have given rise to the *correlation* among the observed objects (genes or microarrays or patients)
- The Principal Components (PC) Model is a straightforward model that seeks to achieve this objective

PCA 3D plot



- Axes represent the first 3 components
- The first 3 components should explain most of the variance
- Formation of clusters
- Relationship of clusters.

Basic Idea of PCA is a Data Reduction Method Based on Analysis of Correlation Pattern(s) That Can Exist Among the Observed Random Variables (i.e. Expression values of Genes).

Raw Data

Array	1	2	...	m
Gene 1	a_{11}	a_{12}	...	a_{1m}
Gene 2	a_{21}	a_{22}	...	a_{2m}
Gene ...	\vdots	\vdots	\vdots	\vdots
Gene n	a_{n1}	a_{n2}	...	a_{nm}

n is the number of genes (gene probes); m is the number of arrays (experiments)

A Structure of Correlation Matrix is the **Major Object for PCA**

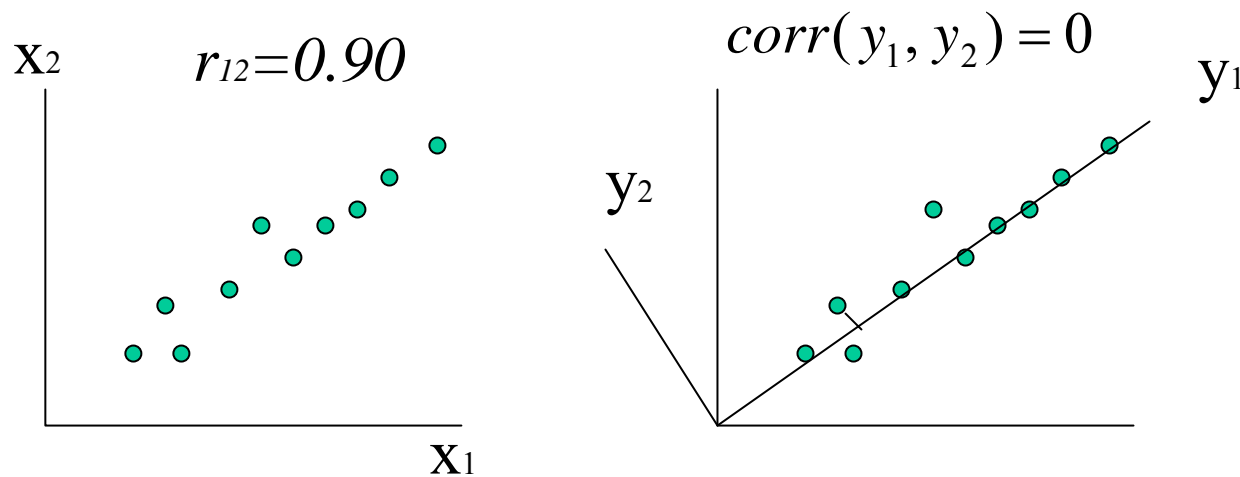
Correlation Matrix	Gene 1	Gene 2	...	Gene n
Gene 1	1	r_{12}	...	r_{1n}
Gene 2	r_{21}	1	...	r_{2n}
Gene ...	\vdots	\vdots	\vdots	\vdots
Gene n	r_{n1}	r_{n2}	...	1

A correlation matrix is a symmetric matrix of correlation coefficients

($-1 \leq r_{ij} \leq 1$ and $r_{ij} = r_{ji}, i, j = 1, 2, \dots, n; r_{ii} = 1$)

The Results of PCA are a small set of the orthogonal (independent) Variables Grouping of the Variables

From a purely mathematical viewpoint the purpose of PCA is to transform n correlated random variables to an orthogonal set which reproduces the original variance/covariance structure.

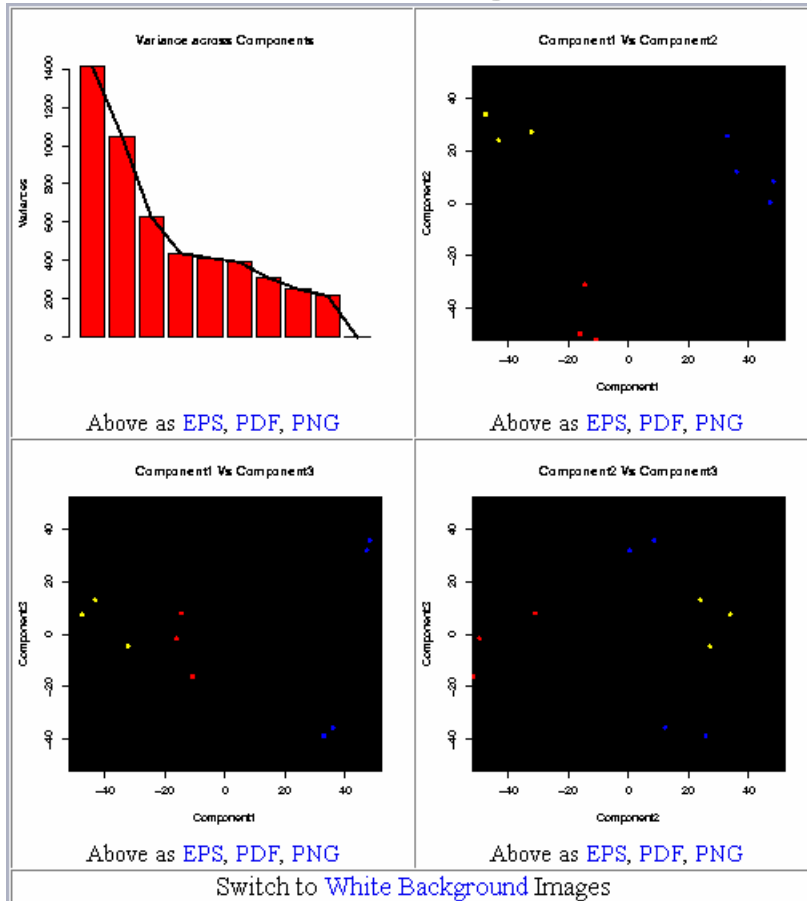


(The First) Principal Component y_1 can “explain” the major fraction ($\sim 90\%$) of a dispersion of variables x_1 and x_2 for all of the 10 observed objects.

Sample: Small Round Blue Cell Tumors (SRBCTs)

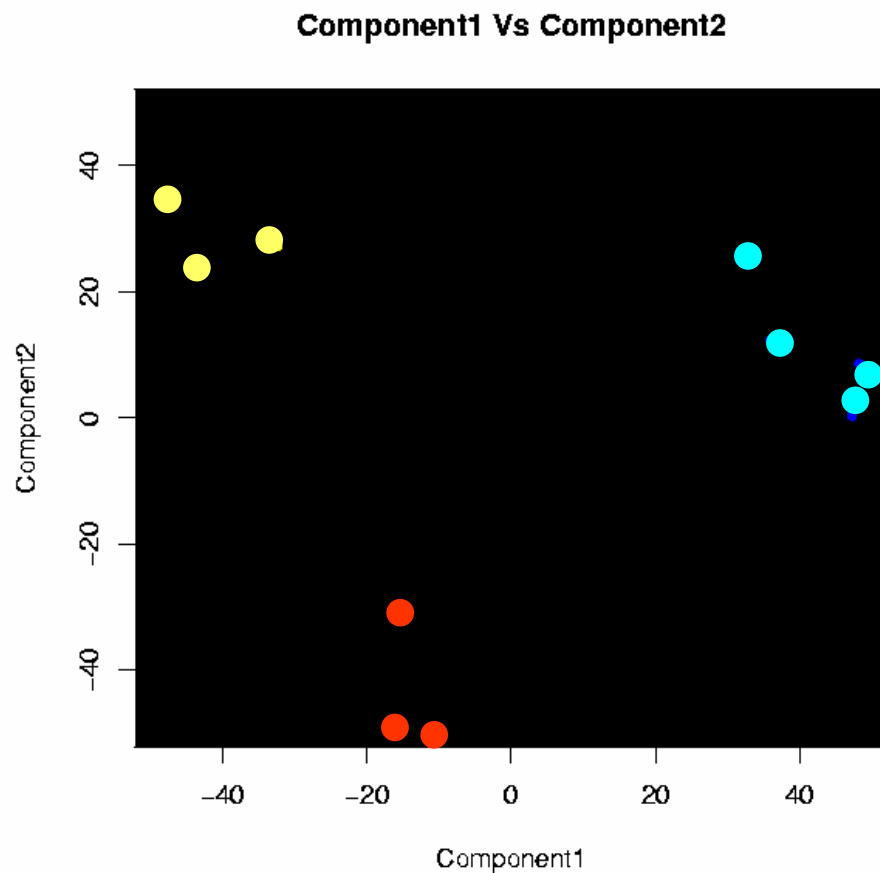
- 63 Arrays representing 4 groups
 - BL (Burkitt Lymphoma, $n_1=8$)
 - EWS (Ewing, $n_2=23$)
 - NB (neuroblastoma, $n_3=12$)
 - RMS (rhabdomyosarcoma, $n_4=20$)
- There are 2308 features (distinct gene probes)

PCA Detailed Plot



- "Scree" plot
- 2-D plots

PCA 2-D plots



- First 2 components separate 3 groups well

MDS overview

(Multidimensional Scaling)

- An alternative for PCA
- Non-linear projection methodology
- Tolerates missing values

Summary of PCA and MDS

- Dimension reduction tools
- Graphic representation to help explain patterns
- Quality control for experimental variance

Hands-on Session 3

- Lab 8
- Total time: 15 minutes
- Next class tomorrow at 1:00 pm

Analyzing Microarray Data using the mAdb System

February 16-17, 2005 1:00 pm - 4:00pm
madb-support@bimas.cit.nih.gov

Day 2

mAdb Analysis Tools

Esther Asaki, Liming Yang, John Powell

Agenda

1. mAdb system overview
2. mAdb dataset overview
3. mAdb analysis tools for dataset
 - Class Discovery - clustering, PCA, MDS
 - Class Comparison - statistical analysis
 - t-test
 - ANOVA
 - Significance Analysis of Microarrays - SAM
 - Class Prediction - PAM

Various Hands-on exercises

Class Comparison

- Statistical distributions of gene expression data
- Hypothesis test and two types of errors
- mAdb statistical analysis tools for class comparison
 - t-test
 - One way ANOVA
 - SAM

Concept of Probability

In Terms of Microarray Data

N: total number of measurements of the expression level of a gene,

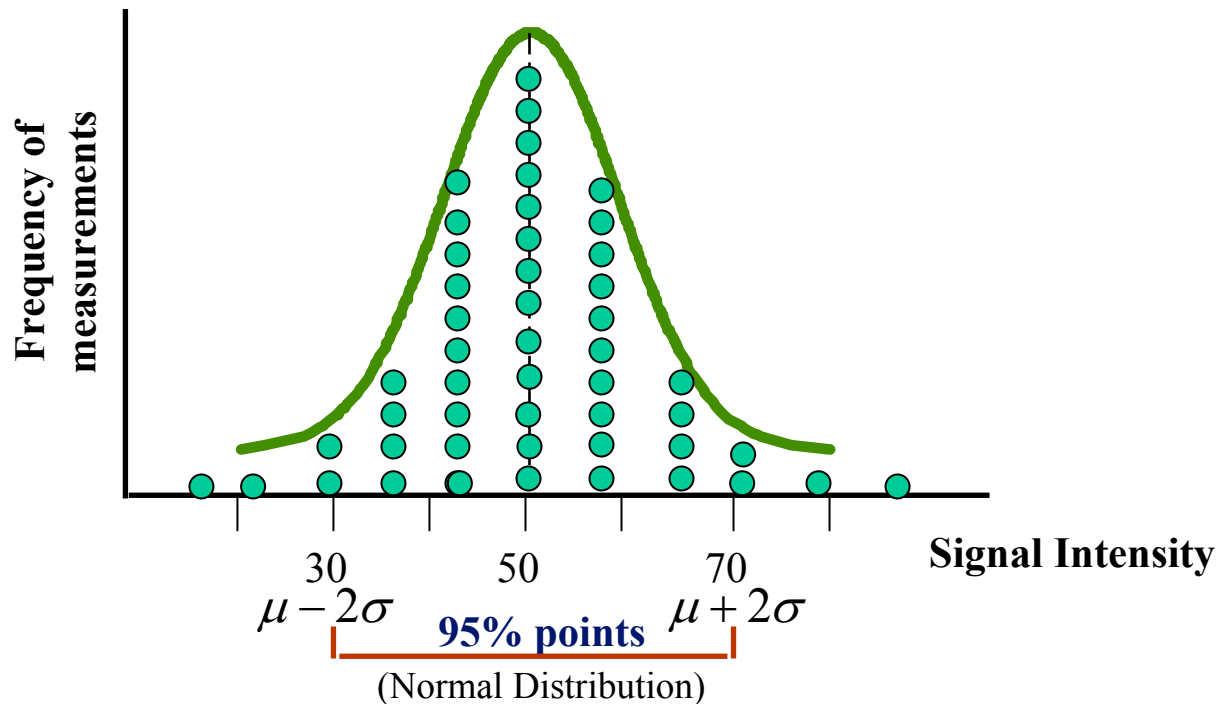
n(A): number of occurrence of expression level A,

Probability: the ratio $n(A)/N$ approximates the probability of getting expression level A, $P(A)$, with increasing accuracy as N increases.

$$P(A) \leftarrow n(A) / N, 0 \leq P(A) \leq 1$$

To get the probabilities of all possible expression levels of a gene, we can construct the **histogram** (the empirical frequency distribution), which approximates the **probability function** of the expression level of that gene.

Replicated Measurements and the Frequency Distribution Function



Center: Mean μ

Spread: Standard deviation σ

Sources of Errors and Uncertainty in Microarray Data Analysis

- Poorly-controlled external factors (quality of tissue sample, RNA etc.)
- Mixture of biological samples derived from many cells and/or complex tissues
- Biological noise (stochastic mechanisms of gene expression)
- Technical noise of background signals
- Limited number of replicates (cost, personnel, etc. constraints)
- Inadequate statistical methods

Technical Caveats

- Technical variability (noise) has a significant intensity bias for low signal intensity values.
- Normalization is required for comparison.
- Simple, static fold change thresholds are too stringent at high intensities and not stringent enough at low intensities.

Statistical and Biological Problems with Fold Change of Means

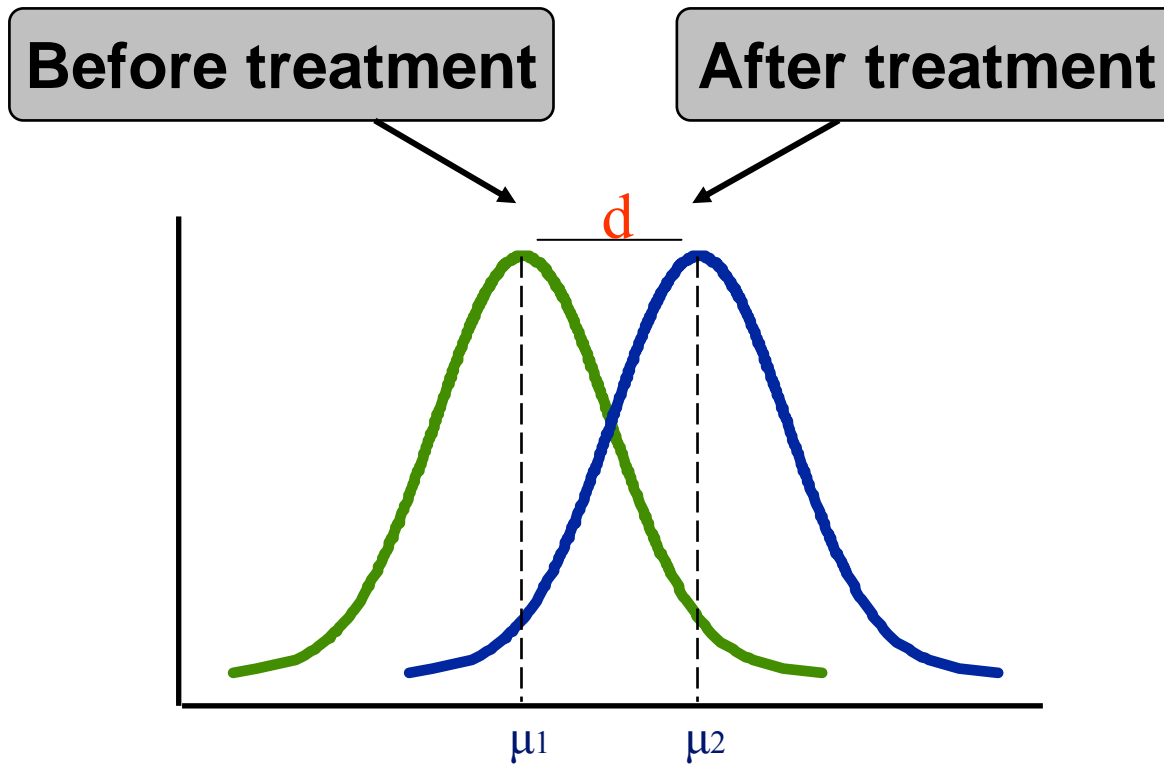
- Genes with high fold change may exhibit high variability among cell types due to natural biological variability for these genes
- Genes with small fold changes may be highly reproducible and should be biologically essential genes

Conclusion

Need robust statistical tests of microarray data

Need additional biological validations

Hypothesis Test



Null hypothesis

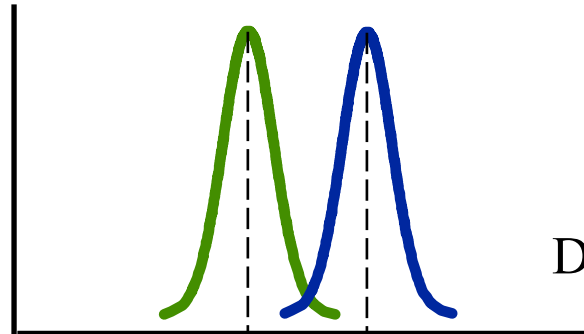
$$H_0 : \mu_1 = \mu_2$$

Alternative hypotheses

$$H_1 : \mu_1 \neq \mu_2$$

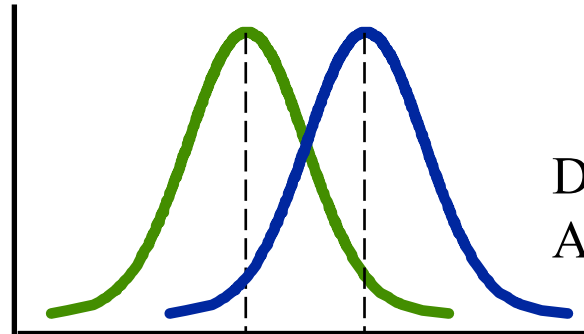
Spread (Variability) of Measurements

**low
variability**



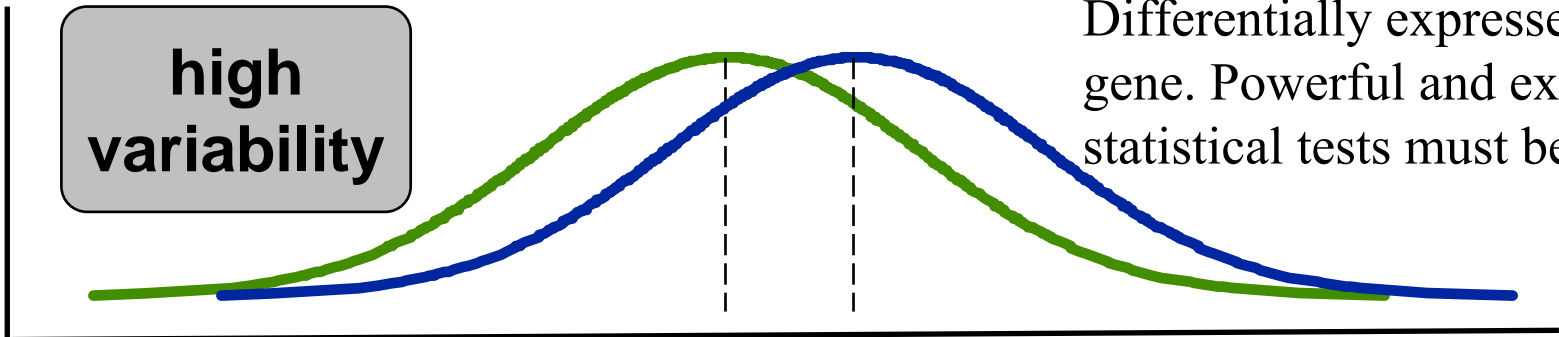
Differentially expressed gene

**medium
variability**



Differentially expressed gene.
A low-reliable estimate

**high
variability**



Differentially expressed
gene. Powerful and exact
statistical tests must be used

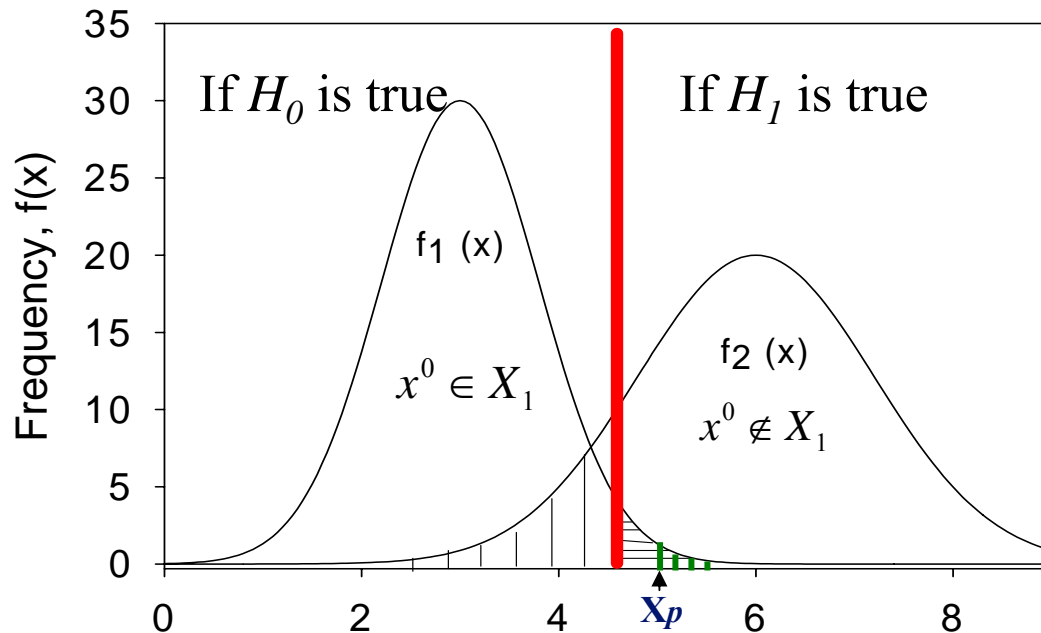
Two Types of Errors

Type I error: Rejecting the null hypothesis while it's true;

Type II error: Accepting the null hypothesis while it's not true.

	Accept H_0	Reject H_0
H_0 is true	Correct decision	Type 1 error False positive
H_0 is false	Type II error False negative	Correct decision

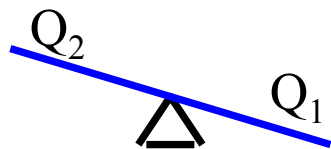
Relation of Type I & Type II Errors



X_1 : data set for control population
 X_2 : data sets for tested population
 x_0 : the critical (the rejection) value of x
 x^o : the observed value of x

Q_1 = The probability of a type I error
 (false-positive)

Q_2 = The probability of a type II error
 (false-negative)



Q_2 x_0 Q_1

Any modifications of x_0 has the opposite effects on probabilities of errors of Type I and Type II: if Q_1 is pushed down, then Q_2 is raised. However, an increase of sample size decreases of both types of errors.

The ***p-value*** is the probability (significance value) at which a true null hypothesis is rejected *by chance only*.

Class Comparison

Goal: To identify differentially expressed genes, i.e. a complete list of genes with expression levels statistically and (more important) biologically different in two or more sets of the representative transcriptomes.

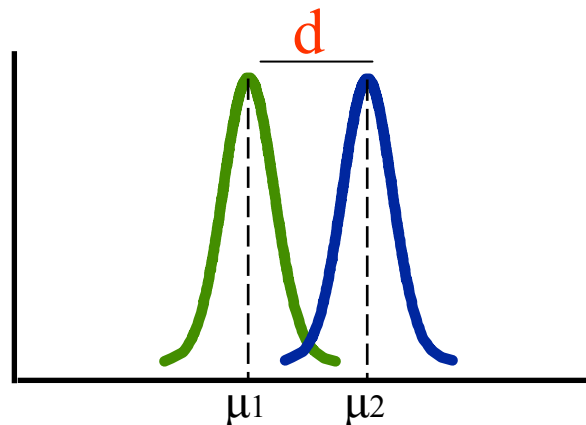
- t-test (2 groups)
- ANOVA (> 2 groups)
- SAM (1, 2, and more groups)

t-Test

The t-test assesses whether the means of two groups are statistically different

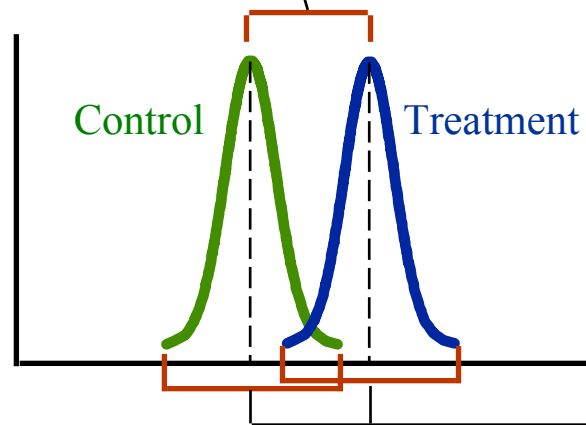
The null hypothesis:

$$H_o : \mu_1 - \mu_2 = 0$$



t-Test (Cont'd)

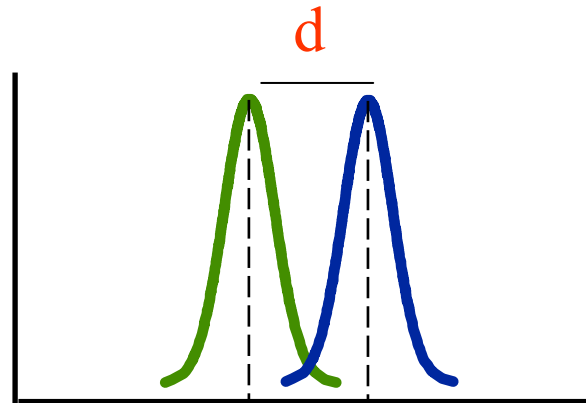
$$\begin{aligned} \frac{\text{signal}}{\text{noise}} &= \frac{\text{difference between group means}}{\text{variability of groups}} \\ &= \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)} \\ &= \text{t-value} \end{aligned}$$



Calculating p-Value (t-Test)

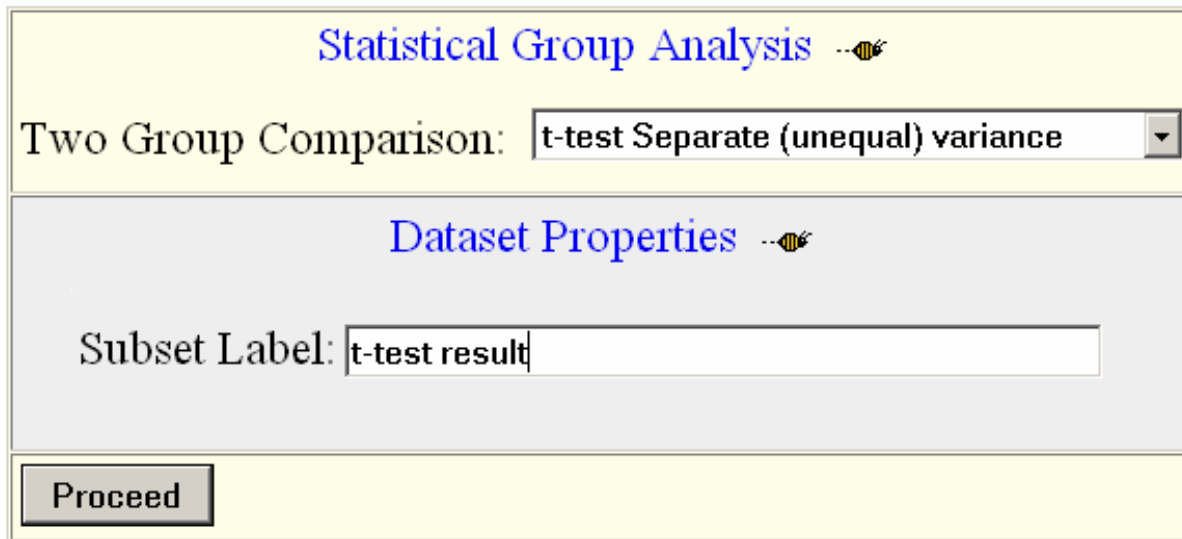
- The p-value is the probability to reject the null hypothesis ($H_o : \mu_1 - \mu_2 = 0$) when it is true (e.g. $p=0.0001$)
- When carrying out a t-test, a p-value can be calculated based on t and the sample sizes n_1 and n_2 .


Large distance d ,
low variability,
large sample sizes,
then small p ,
i.e. more significant.






mAdb t-Test

- 2 group statistic analysis automatically selected for a 2 group dataset



Statistical Group Analysis 

Two Group Comparison: t-test Separate (unequal) variance  

Dataset Properties 

Subset Label: t-test result

Proceed

t-Test Results

						$\log_2(A) - \log_2(B)$	
A	A	A	B	B	B	↓ ↑	↓ ↑
JIM3_A	JJN3_A	U266_A	HDLM2_A	L428_A	L540_A	p-Value	Difference
52.4309	54.9520	45.0046	0.7800	0.6485	0.8532	1.9737e-06	6.07
35.1142	52.4541	42.8235	0.7800	0.6485	0.8532	8.9006e-06	5.83
53.3166	74.5535	46.5118	0.7800	0.6485	0.8532	1.1662e-05	6.24
5.9693	5.9444	5.7954	9.4782	9.6511	10.0555	1.4619e-05	-0.72
12.2739	13.0063	9.6026	0.7800	0.6485	0.8532	2.4704e-05	3.93
0.6680	0.6954	0.6536	9.0445	8.4780	13.0657	3.7853e-05	-3.9
3.7943	3.4277	3.3739	7.3190	7.6012	7.2551	4.7738e-05	-1.07
0.6680	0.6954	0.6536	2.3401	2.0402	2.5358	4.9127e-05	-1.77
0.6680	0.6954	0.6536	7.6466	6.0506	9.6493	5.7477e-05	-3.51
0.6680	0.6954	0.9490	8.0788	8.5636	6.8106	5.8369e-05	-3.35
0.6680	0.6954	0.7869	68.9017	34.0804	72.9403	6.3509e-05	-6.28
34.7315	29.5014	60.8882	0.7800	0.6485	0.8532	7.1258e-05	5.71
0.6680	0.6954	0.6706	0.8424	0.8593	0.8532	8.4299e-05	-0.329
0.6680	0.6954	0.6536	39.1841	17.6407	27.2176	9.1539e-05	-5.31
3.7288	2.9875	3.1098	0.9774	0.8392	0.8532	9.9425e-05	1.88
0.6680	1.3275	0.6536	26.2949	22.3119	26.9078	0.00014347	-4.91
1.7328	1.8435	2.0412	0.8557	0.9196	0.8532	0.00014599	1.09

Statistic Results Filtering

Check boxes on the left to activate specific filters
▼

<input checked="" type="checkbox"/>	T-test p-value (two tailed)	<input "="" type="text" value="<="/>	<input type="text" value="0.001"/>
<input checked="" type="checkbox"/>	Group mean Difference	<input "="" type="text" value=">="/>	<input type="text" value="1"/>
<input checked="" type="checkbox"/>	Apply <i>Symmetrically</i>		


Subset Label:

(Optional)

← statistical significance,
i.e. p-value

← $\overline{\log_2(A)} - \overline{\log_2(B)}$

Other Statistical Tests for 2 Group Comparison

Statistical Group Analysis 

Two Group Comparison:

Dataset P

Subset Label:

- Select a Method
- Paired t-test
- t-test Pooled (equal) variance
- t-test Separate (unequal) variance
- Wilcoxon Rank-Sum (Mann Whitney U)
- Wilcoxon Matched-Pairs Signed-Rank

Parametric (normal distribution)

Non-Parametric (distribution free)

Multiple Group Comparison

	Group 1	Group 2	...	Group k
Gene 1	$\mu_{1.1}$	$\mu_{1.2}$...	$\mu_{1.k}$
Gene 2	$\mu_{2.1}$	$\mu_{2.2}$...	$\mu_{2.k}$
...
Gene n	$\mu_{n.1}$	$\mu_{n.2}$...	$\mu_{n.k}$

n: Number of genes/probes

k: number of groups, $k > 2$

Analysis of Variances (ANOVA)

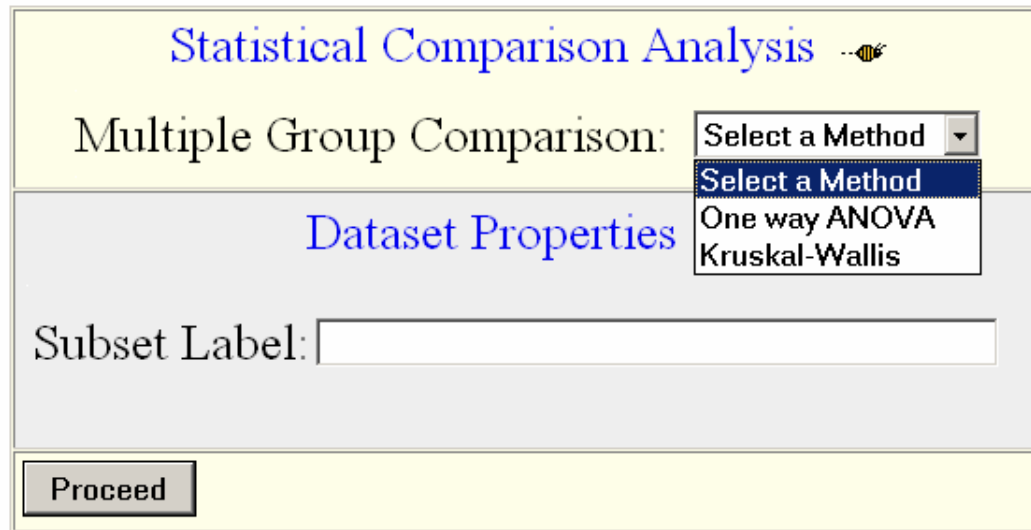
To compare several population means:

$$H_o : \mu_1 = \mu_2 = \dots = \mu_k \quad (k > 2)$$

vs.

$$H_1 : \mu_i \neq \mu_j; \quad \text{for some } 1 \leq i \neq j \leq k$$







Multiple Group Comparison



The image shows a software dialog box titled "Statistical Comparison Analysis" with a small icon to the right. Inside the dialog, the text "Multiple Group Comparison:" is followed by a dropdown menu. The dropdown menu is open, showing three options: "Select a Method" (highlighted in blue), "One way ANOVA", and "Kruskal-Wallis". Below the dropdown, the text "Dataset Properties" is displayed. Underneath, there is a label "Subset Label:" followed by an empty text input field. At the bottom left of the dialog, there is a button labeled "Proceed".

- ANOVA: parametric test based on F-statistics
- Kruskal-Wallis : non-parametric rank-based test

ANOVA Results and Filtering

  p-Value	  Difference	  Groups
9.6276e-22	4.11	A-B
3.488e-20	2.99	D-C
2.5008e-19	3.59	A-B
2.5733e-18	2.59	A-D
1.4459e-17	2.76	D-A
5.7703e-17	2.89	A-B
8.728e-17	3.14	D-B
1.3957e-16	3.95	C-A
4.1114e-16	4.03	A-B
1.4464e-15	3.76	A-B
2.369e-15	3.1	D-B
7.4515e-15	3.32	A-B
8.187e-15	2.76	A-C
2.5078e-14	4.1	A-B
2.5526e-14	5.68	D-B

← Group Pair for Max Mean Difference



Maximum Difference between Group Means

Check boxes on the left to activate specific filters

▼

☒ One Way ANOVA p-value <=

☒ Group mean Difference >=

Subset Label:
(Optional)

Hands-on Session 4

- Lab 9
- Total time: 15 minutes

Multiple Testing

- Large-scale experiments & statistical problems
 - Finding the differentially expressed genes measured simultaneously in the two or more groups of microarrays has a problem of multiple comparison, where many null hypotheses are tested simultaneously.
- p-value adjustment
 - Although p-value cut off (α) of 0.01 is significant in a conventional single-variable test, a microarray experiment for 20,000 gene probes would identify $20,000 \times 0.01 = 200$ genes just by chance!

Correcting Multiple Testing

- False Discovery Rate (FDR)
- Significance Analysis of Microarrays (SAM):
Modified t-test for selection of the differentially expressed gene sets, and calculates FDR
- <http://www-stat.stanford.edu/~tibs/SAM/index.html>
- Estimation of the parameters of the model and their optimization
- Interpretation of SAM output and SAM plot

Multiple Testing and FDR

	Not rejected	Rejected	Total
H_0 True	$m_0 - V$	V	m_0
H_1 True	$m_1 - S$	S	m_1
Total	$m - R$	R	m

m : # hypothesis
 V : # false positive
 R : # significant hypothesis

Probability of false-positive gene discovery:

$$\text{False Discovery Rate (FDR)} = E(V/R \mid R > 0) * \Pr(R)$$

Significance Analysis of Microarrays (SAM)

- To select a fairly large number of differentially expressed genes, accepting some *falsely significant* genes, as long as their number is relatively small compared to the total number of significant genes, selected at significance level α .
- For one or two groups, SAM computes a t-like statistic $d(i)$ for each probe i ($i=1,2,\dots,n$), measuring the difference between the normalized mean signals of the groups.
- For more groups, SAM computes a F-like statistic.

SAM for 2 groups

I. The “relative difference” $d(i)$ in gene expression for two groups I and U of repeated samples is:

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

$\bar{x}_I(i)$: average expression level for gene i in group I,

$\bar{x}_U(i)$: average expression level for gene i in group U,

$s(i)$: the estimate of standard deviation of repeated measurements,

s_0 : the fudge factor that reduces the “relative differences” of the low expressed genes (noise) and/or genes with similar expression levels. i.e. $d(i)$ will not be too large with small $s(i)$.

Permutations of the Arrays and the Expected Relative Differences

Group I Group U

a1	b1
a2	b2
a3	b3
a4	b4

Group I Group U

b1	a1
a2	b2
a3	b3
a4	b4

Group I Group U

b1	a1
a2	b2
b3	a3
a4	b4

n: the number of hybridized signals (gene probes);

k: the number of permutations of arrays between the groups.

Permutation 1: $d_1(1) \leq \dots \leq d_1(n)$

.....

Permutation p: $d_p(1) \leq \dots \leq d_p(n)$

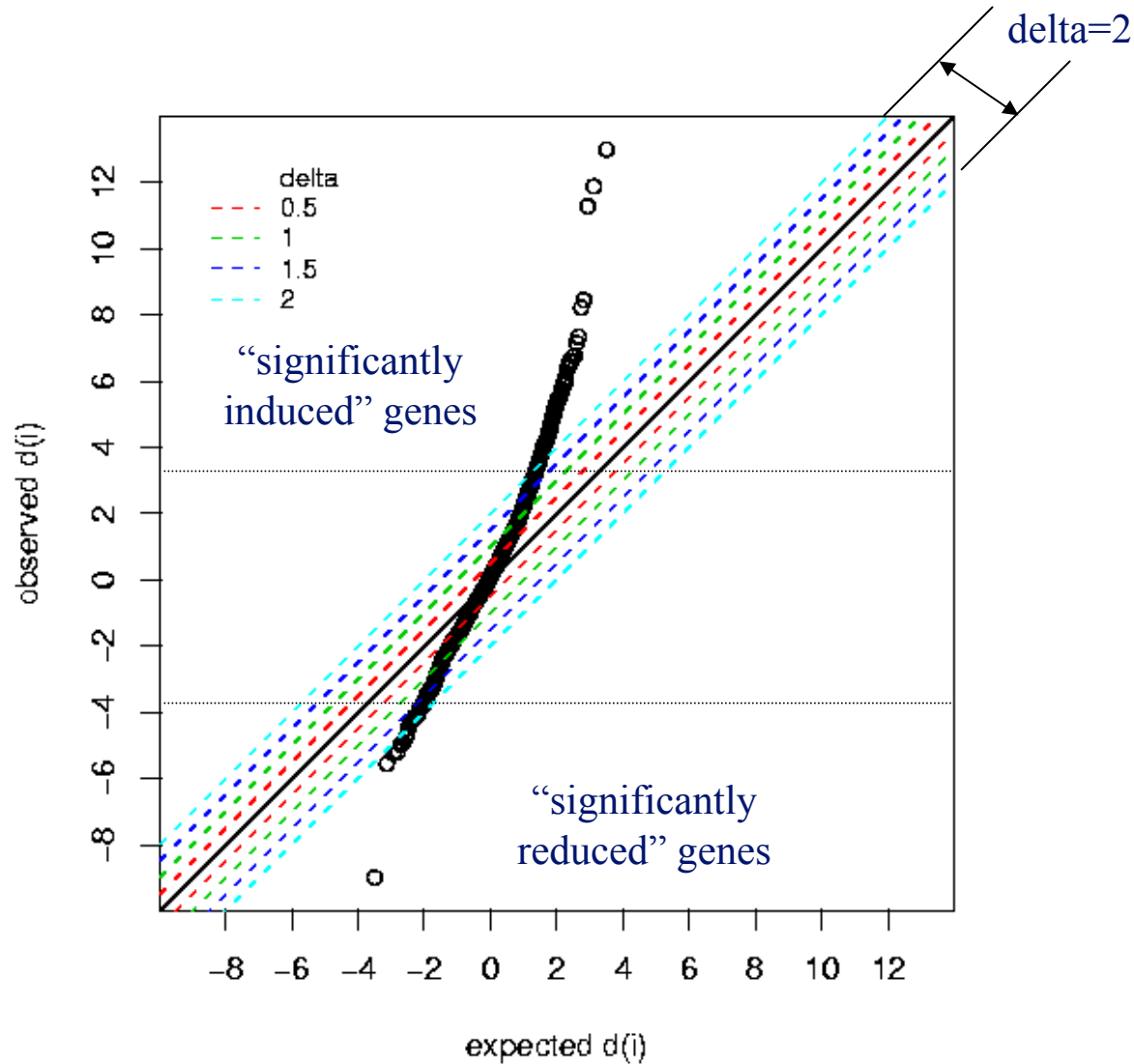
.....

Permutation k: $d_k(1) \leq \dots \leq d_k(n)$

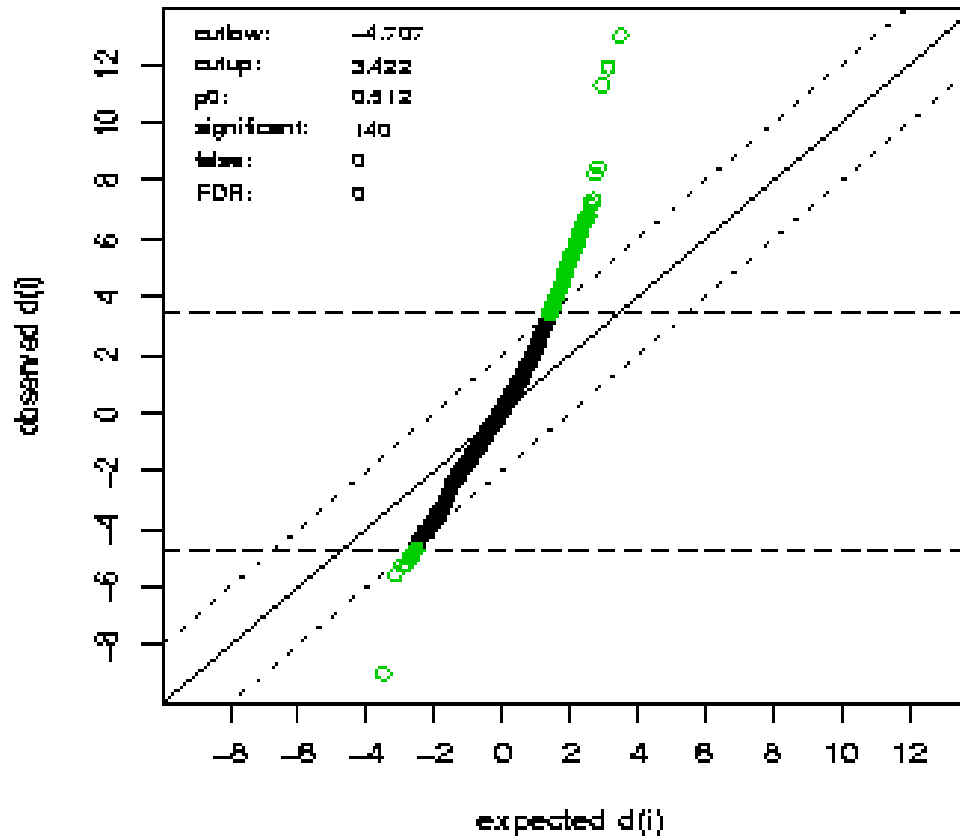
$$\bar{d}(i) = \frac{1}{k} \sum_{i=1}^k d_p(i)$$

Expected relative difference
for gene i ($i=1,2,\dots,n$)

SAM Plot for a Set of Delta



SAM Plot for Delta = 2



FDR estimation:

The ratio of average number of falsely significant gene probes exceeding a given cutoff value (delta) to the number of significance gene probes at that cutoff₁₀₃

mAdb SAM

from Dataset: **Small, Round Blue Cell Tumors (SRBCTs)**, Nature Medicine Vol 7, Num 6, 601-673 (2001)

Interactive Array Filtering

88 arrays and 2308 genes in the original data set

11 arrays and 2308 genes in the output data set.

77 arrays were excluded by interactive grouping/filtering

No arrays were restored by interactive grouping/filtering

6 arrays assigned to Group A

5 arrays assigned to Group B

View the complete [History](#).

[Expand](#) this Dataset.

Access Datasets in your [Temporary](#) area.

Filtering/Grouping/Analysis Tools

Choose a Tool **BETA SAM: Significance Analysis of MicroArrays** and **Proceed**

Interactive Graphical Viewers

Choose a Viewer **MDS: MultiDimensional Scaling** and **View**

Dataset Retrieval & Display Options

Retrieve Dataset formatted for **Eisen Cluster**

Redisplay

☐ Show Array Details at the top of the page

Background Color **- None -** Contrast **1.568**

Limiting display to **to 25 genes**

☒ Show Data Values ☒ Use Names in Column Heading

☐ Apply log2 transform ☒ Use Description in Column Heading

☒ Show Gene Symbols ☒ Show Map Information

SAM Data

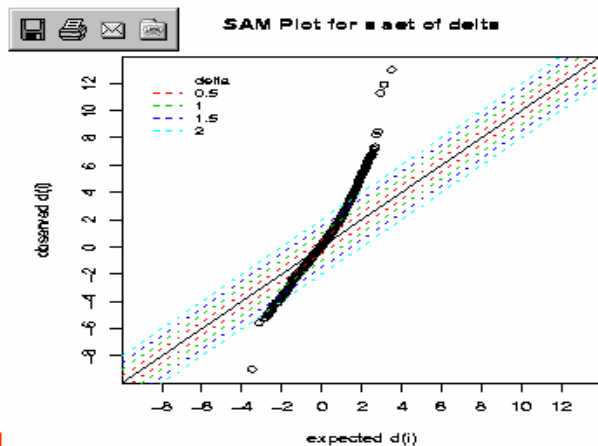
Records 1 to 25 of 2308 total records displayed.

A	A	A	A	A	A	B	B	B	B	B	<div><div>⬇</div><div>⬆</div><div>⬇</div><div>⬆</div></div>			
EWS-T1	EWS-T2	EWS-T3	EWS-T4	EWS-T6	RMS-C10	BL-C5	BL-C6	BL-C7	BL-C8	BL-C1	Aver	Mx-Mn	Well ID	Feature J
3.2025	1.6547	3.2779	1.0060	2.7098	1.5410	0.2989	0.1856	0.1045	0.3178	0.1437	-0.5176	4.971	1080460	IMAGE:21
0.0681	0.0710	0.1160	0.1906	0.2367	0.0672	0.0839	0.1283	0.0994	0.0494	0.0563	-3.4113	2.260	1080461	IMAGE:25
1.0460	1.0409	0.8926	0.4302	0.3693	0.6689	1.0989	1.7574	0.2362	0.9711	1.0739	-0.3952	2.895	1080462	IMAGE:26
0.1243	0.0520	0.1014	0.1035	0.2190	1.1397	1.3145	1.3695	1.2625	1.2685	0.1198	-1.5803	4.719	1080463	IMAGE:22
0.4941	0.2045	0.2818	0.2984	0.3711	0.0401	0.3285	0.1284	0.1687	0.0573	0.3935	-2.3232	3.623	1080464	IMAGE:22
3.1207	2.1609	1.9773	1.6804	1.7800	0.5750	0.7530	0.5325	0.9698	1.0432	2.3396	0.4040	2.551	1080465	IMAGE:22
3.7106	2.4452	3.2590	5.8901	3.2376	6.1087	3.0222	4.8113	4.6305	3.7375	3.3334	1.9511	1.321	1080466	IMAGE:23
1.8416	1.1473	1.4106	0.2958	0.6769	0.5264	2.2284	1.1472	0.6647	0.5825	1.0947	-0.1413	2.913	1080467	IMAGE:23
1.2607	0.7371	0.9548	0.7381	0.8546	0.7117	1.4646	2.8207	2.2148	1.2009	2.2681	0.3010	1.987	1080468	IMAGE:24
2.9001	1.9989	2.0775	1.6610	0.6808	1.0343	2.0438	2.6476	1.4568	1.6544	1.8761	0.7663	2.091	1080469	IMAGE:25
4.0270	2.6131	4.8139	4.9105	4.5104	7.6119	4.3938	4.5243	5.8249	5.6817	4.6666	2.2404	1.542	1080470	IMAGE:31
1.0643	0.8541	0.4257	1.5866	0.6461	0.6720	0.5792	0.7810	0.8217	0.8692	1.3787	-0.2795	1.898	1080471	IMAGE:31
4.0651	4.7284	4.7120	9.4802	3.6433	7.4534	4.1207	5.1979	4.6583	6.1956	4.8565	2.3692	1.380	1080472	IMAGE:27
1.4730	2.4784	2.7548	0.1667	1.7957	1.1213	1.9109	1.5108	1.2601	0.5194	0.7753	0.2096	4.047	1080473	IMAGE:27
2.7932	1.5103	1.9162	1.1314	1.0375	0.2976	0.6045	0.5331	0.4699	0.5765	0.9937	-0.1845	3.230	1080474	IMAGE:27
0.4815	0.8961	1.2710	2.6361	0.3976	0.1707	0.3801	0.4887	0.3888	0.4765	0.2920	-0.9078	3.949	1080475	IMAGE:28
1.4482	0.4850	1.1331	0.8405	0.8846	1.0015	0.8015	0.9010	0.6565	0.6765	0.5780	-0.2897	1.578	1080476	IMAGE:29
3.3214	2.3431	2.4818	0.9928	1.5156	1.1450	1.4196	1.6072	2.7400	2.5685	3.7927	0.9986	1.934	1080477	IMAGE:34
0.7022	0.2531	2.0350	0.1239	1.2582	1.5526	0.9754	1.9857	2.0390	0.8393	0.6224	-0.2245	4.041	1080478	IMAGE:35
1.7260	1.7841	1.7340	0.5216	1.0114	0.3846	0.8591	1.2435	1.0625	0.5903	1.9728	0.0471	2.359	1080479	IMAGE:32
1.5136	1.0886	2.6863	0.9867	1.5428	2.0073	0.8364	0.8401	1.0302	0.4571	1.4035	0.2410	2.555	1080480	IMAGE:32
3.9255	5.9544	5.5842	4.8170	5.1313	6.1289	5.3582	4.2545	6.3473	6.3606	5.0796	2.4056	0.696	1080481	IMAGE:33
0.5296	0.5337	1.1332	0.6451	0.6248	0.4066	0.3915	0.4413	0.2791	0.1034	0.5852	-1.1521	3.454	1080482	IMAGE:33
3.9098	2.7007	4.6055	2.0627	4.4183	6.0772	3.4114	3.9238	5.0894	4.3511	4.7345	1.9864	1.559	1080483	IMAGE:34
3.7136	3.2339	2.2437	3.1900	2.1173	2.9153	4.1731	2.5854	4.0399	4.4925	4.8857	1.7230	1.206	1080484	IMAGE:34

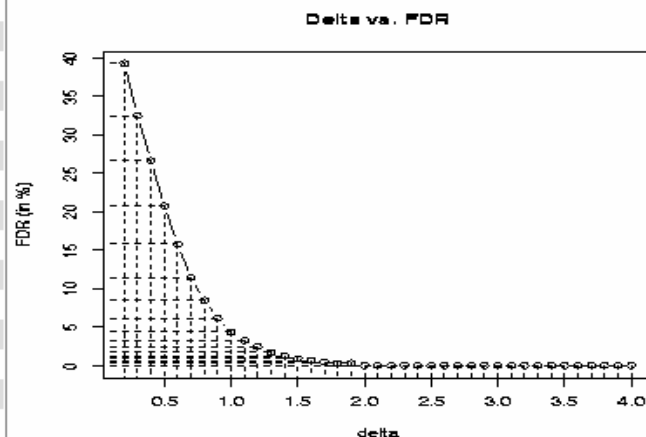
mAdb SAM Results I

Clicking on a Delta value creates a new data Subset or
 ▼ a Delta value at the bottom and Click "Create Subset".

Delta	# of Sig. Genes	# of False Genes	FDR
0.200	1815	1395	0.3934
0.300	1542	979	0.3250
0.400	1412	736	0.2670
0.500	1208	490	0.2078
0.600	1052	325	0.1581
0.700	849	189	0.1142
0.800	710	118	0.0851
0.900	587	70	0.0615
1.000	492	42	0.0437
1.100	448	28	0.0326
1.200	396	19	0.0246
1.300	359	12	0.0171
1.400	328	8	0.0125
1.500	283	5	0.0090
1.600	233	3	0.0066
1.700	220	2	0.0047
1.800	186	1	0.0028
1.900	156	1	0.0033
2.000	140	0	0.0000
2.100	120	0	0.0000
2.200	115	0	0.0000
2.300	98	0	0.0000
2.400	87	0	0.0000
2.500	74	0	0.0000
2.600	62	0	0.0000
2.700	57	0	0.0000
2.800	52	0	0.0000
2.900	49	0	0.0000
3.000	45	0	0.0000
3.100	42	0	0.0000
3.200	38	0	0.0000
3.300	33	0	0.0000
3.400	32	0	0.0000
3.500	25	0	0.0000
3.600	21	0	0.0000
3.700	10	0	0.0000



Above as [EPS](#), [PDF](#), [PNG](#)



Above as [EPS](#), [PDF](#), [PNG](#)



mAdb SAM Results II

SAM d statistics
(normalized distance)

Significance value
(lowest FDR)

Records 1 to 25 of 283 total records displayed.

A	A	A	A	A	A	B	B	B	B	B						
EWS-T1	EWS-T2	EWS-T3	EWS-T4	EWS-T6	RMS-C10	BL-C5	BL-C6	BL-C7	BL-C8	BL-C1	Aver	Mx-Mn	d(i)	s(i)	q-value	R.difference
0.3945	0.2986	0.2757	0.2607	0.4046	0.4103	2.1557	1.9916	1.7072	1.3449	2.4256	-0.4450	3.218	-8.9834	0.1900	0.00077	-2.4950
0.5752	0.4646	0.5121	0.2998	0.2031	1.1779	2.6250	3.2456	4.0725	2.5397	3.2127	0.1340	4.326	-5.5538	0.4058	0.00171	-2.7408
0.4147	0.2785	0.2982	0.5336	0.3906	0.2064	2.0555	1.5561	3.1465	1.3678	0.8978	-0.5247	3.930	-5.2410	0.3496	0.00233	-2.2919
0.5677	0.4885	0.6398	0.6545	0.4830	0.4298	1.3266	2.1554	1.4731	1.3334	1.0776	-0.2524	2.326	-5.1946	0.1848	0.00233	-1.4152
0.2151	0.2423	0.3112	0.1954	0.3141	0.1790	0.7768	0.4494	0.7267	0.8581	1.1068	-1.3189	2.628	-4.9903	0.2460	0.00277	-1.6651
0.3098	0.3648	0.3268	0.6515	0.4455	0.5171	1.1521	1.4527	1.4889	1.6262	0.8665	-0.5166	2.392	-4.9316	0.2391	0.00278	-1.6116
0.1844	0.2833	0.2713	0.3253	0.2585	0.0756	0.7904	0.9508	1.3086	1.3896	0.7992	-1.2101	4.200	-4.8040	0.3838	0.00315	-2.2651
0.6507	0.0840	0.1838	0.3586	0.3173	0.4342	1.1608	1.6185	3.7828	2.5348	3.2213	-0.4652	5.493	-4.7068	0.5499	0.00327	-3.0012
0.0860	0.1757	0.1148	0.1902	0.2746	0.0852	1.5769	1.2605	2.0804	1.9797	0.2545	-1.4477	4.610	-4.4672	0.5933	0.00476	-3.0423
0.3246	0.3545	0.2067	0.0578	0.2635	0.1197	0.9197	1.1204	0.9938	1.0349	0.8334	-1.3339	4.277	-4.3639	0.4576	0.00515	-2.3797
0.4628	0.5709	0.4914	0.7941	0.3252	0.3888	1.7641	1.2860	1.5374	1.2258	0.9175	-0.3900	2.440	-4.2811	0.2482	0.00568	-1.4379
1.5115	1.5096	0.8966	1.2869	0.7250	0.7268	2.7377	3.1414	2.2796	2.6268	3.4603	0.7227	2.255	-4.2585	0.2450	0.00578	-1.4168
0.4020	0.2550	0.3152	0.3846	0.2843	0.1679	0.4970	0.8778	0.8845	1.3243	1.1308	-1.0464	2.980	-4.1966	0.3003	0.00602	-1.6283
0.1647	0.0735	0.1826	0.1091	0.1377	0.0525	0.6268	0.3213	0.3043	0.5631	0.3504	-2.3181	3.578	-4.1629	0.3722	0.00602	-1.9143
0.3863	0.6338	0.3197	0.1897	0.2734	0.0423	3.7478	2.2453	1.8450	1.3306	1.4104	-0.6948	6.469	-4.1253	0.6535	0.00602	-3.0575
0.3109	0.6175	0.3248	0.1131	0.4431	0.1240	1.1211	1.0641	1.4479	1.9277	4.4673	-0.6763	5.304	-4.1116	0.5612	0.00602	-2.6679
0.3257	0.8323	0.4872	0.2898	0.7987	3.9015	5.5857	4.5178	5.6381	5.8817	4.8603	0.8111	4.343	-4.1068	0.6207	0.00602	-2.9092
0.3721	0.4905	0.3745	0.2239	0.1621	0.4997	0.7104	1.2036	1.0815	1.1202	1.1465	-0.8553	2.892	-4.0817	0.3188	0.00602	-1.6589
0.5043	0.8385	0.5527	0.8775	0.3906	0.5454	1.5849	1.4227	2.2138	1.7559	1.0650	-0.1176	2.503	-4.0574	0.2567	0.00619	-1.3973
0.4160	0.6563	0.5589	0.3437	0.3238	0.3107	0.8792	0.9426	1.4504	2.2569	3.2467	-0.4025	3.385	-4.0029	0.3842	0.00655	-1.8888
0.7256	0.7514	0.5969	0.3064	0.4322	0.9539	2.0728	1.4237	1.2331	1.8964	2.0909	-0.0685	2.771	-3.9296	0.3036	0.0071	-1.5376
0.2171	0.1918	0.2113	0.1568	0.3634	0.1631	0.6020	0.4455	0.8289	0.5404	0.3778	-1.6400	2.402	-3.9028	0.2635	0.0071	-1.3707
0.5048	0.1327	0.5114	0.2240	0.6995	0.0670	4.8524	5.0288	4.1432	1.8065	0.7500	-0.3949	6.230	-3.9009	0.7650	0.0071	-3.3264
0.7339	0.9880	0.3495	0.3034	0.3344	0.2990	1.4124	1.1339	2.0510	1.2561	2.1298	-0.3534	2.833	-3.8825	0.3753	0.00712	-1.7976
0.7128	0.8795	0.1936	0.2207	0.1427	0.9624	3.5737	2.2461	2.4018	2.3751	1.5855	-0.1732	4.646	-3.8567	0.5823	0.00748	-2.5841

Hands-on Session 5

- Lab 10
- Total time: 15 minutes

3. mAdb dataset analysis tools

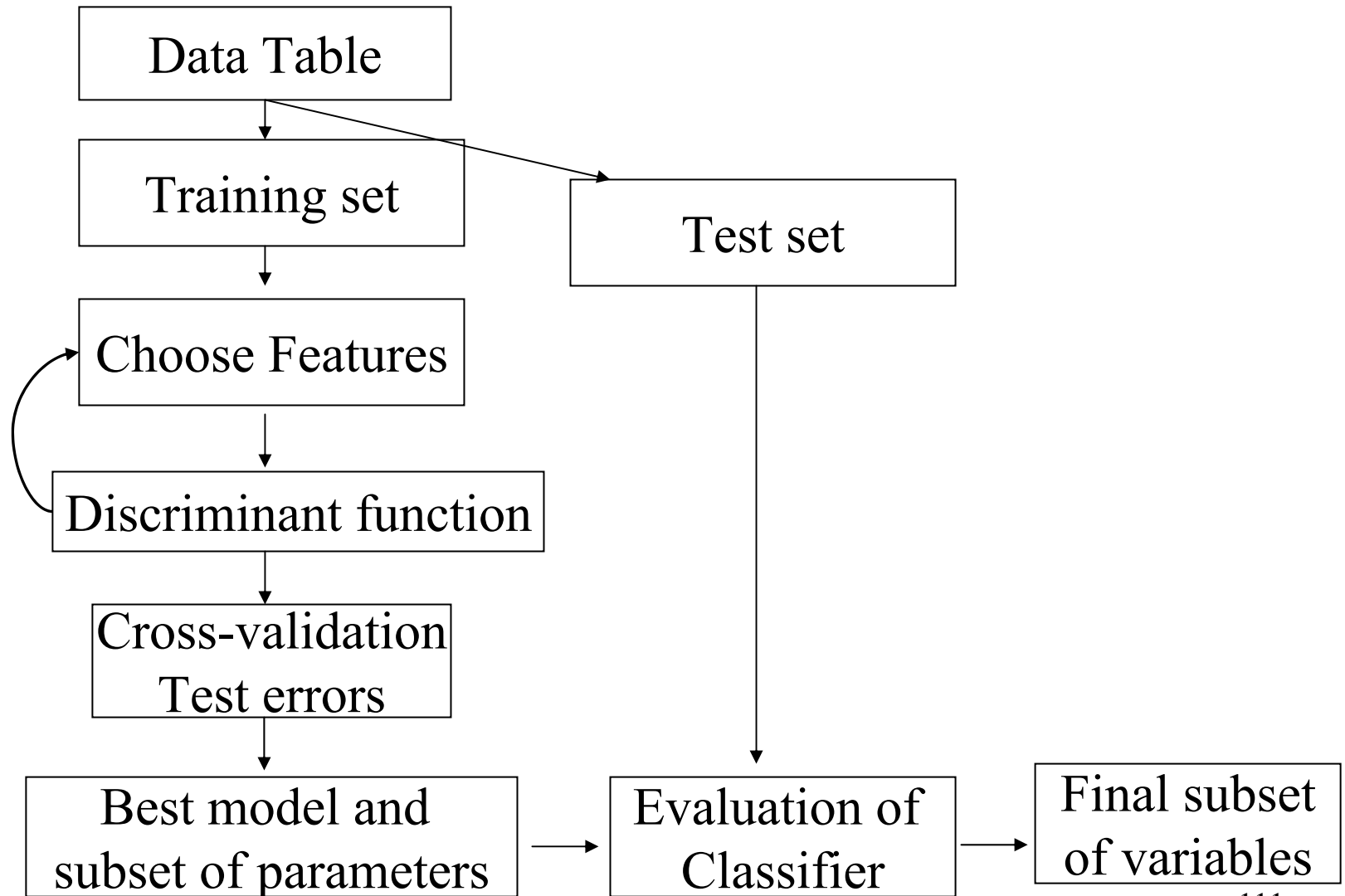
- Class Discovery: clustering, PCA, MDS
- Class Comparison: statistical analysis
- Class Prediction: PAM

Class Prediction

Supervised Model for Two or More Classes

- Prediction Analysis for Microarrays (PAM)
- <http://www-stat.stanford.edu/~tibs/PAM>
- Provides a list of significant genes whose expression characterizes each class
- Estimates prediction error via cross-validation
- Imputes missing values in dataset

Design of the PAM algorithm



Calculating the Discriminant Function

For each gene, a centroid (sample mean) is calculated for each class.

Standardized centroid distance are calculated:

the average gene expression value in each class minus the overall gene expression average value, divided by the standard deviation-like normalization factor (NF) for that gene.

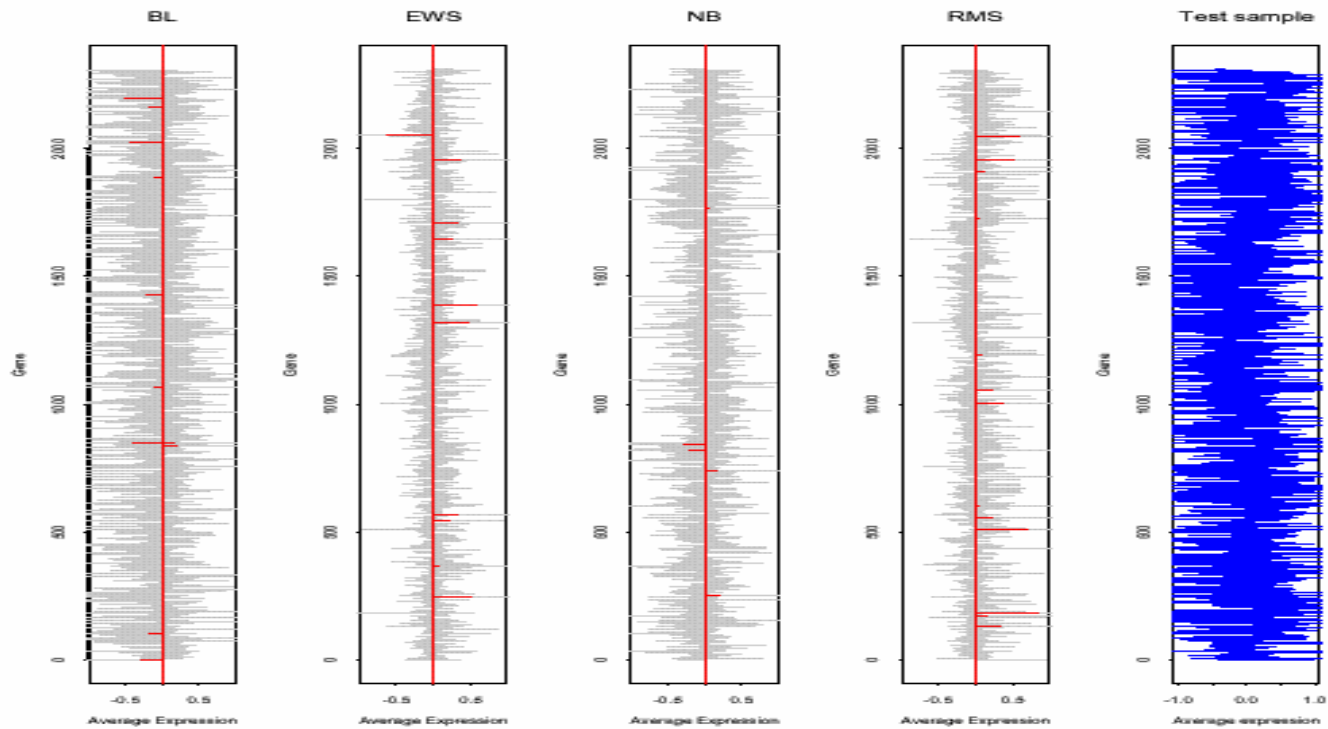
$$\text{centroid distance} = (\text{class avg} - \text{overall avg}) / \text{NF}$$

Creates a normalized average gene expression profile for each class

Class Centroids

SL&DM ©Hastie & Tibshirani March 26, 2002 Supervised Learning: 31

Class centroids

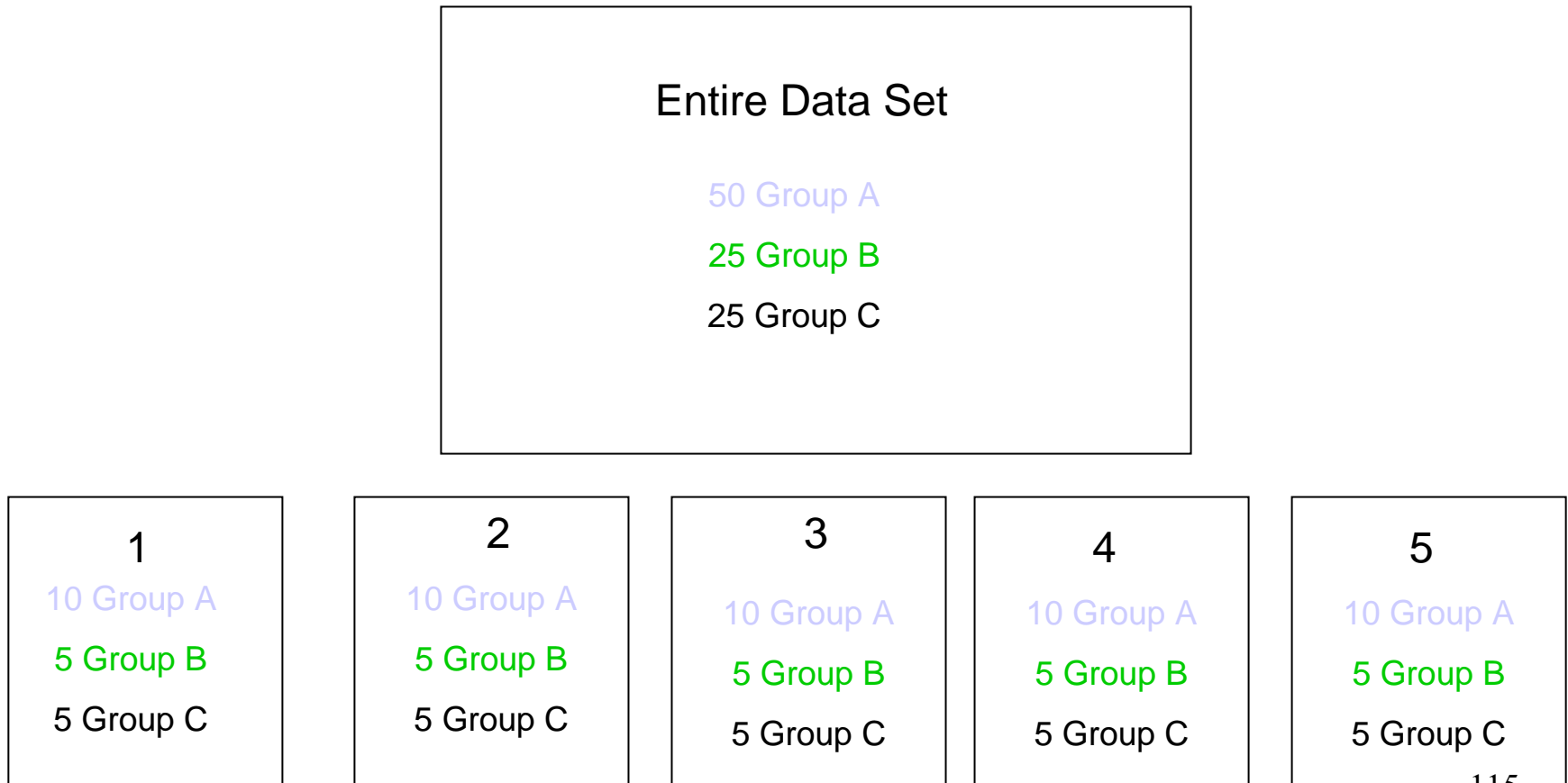


Classifying an Unknown Sample

A classifier takes the gene expression profile of a new sample (microarray) from test sets, and compares it to each of these class centroids. The class whose centroid that it is closest to, in squared distance, is the predicted class for that new sample.

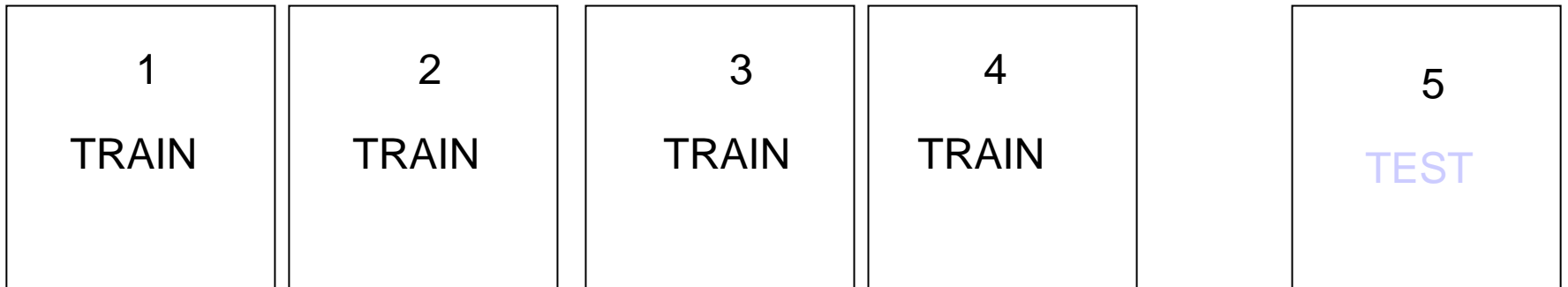
K-fold Cross Validation

- The samples are divided up at random into K roughly equally sized parts.

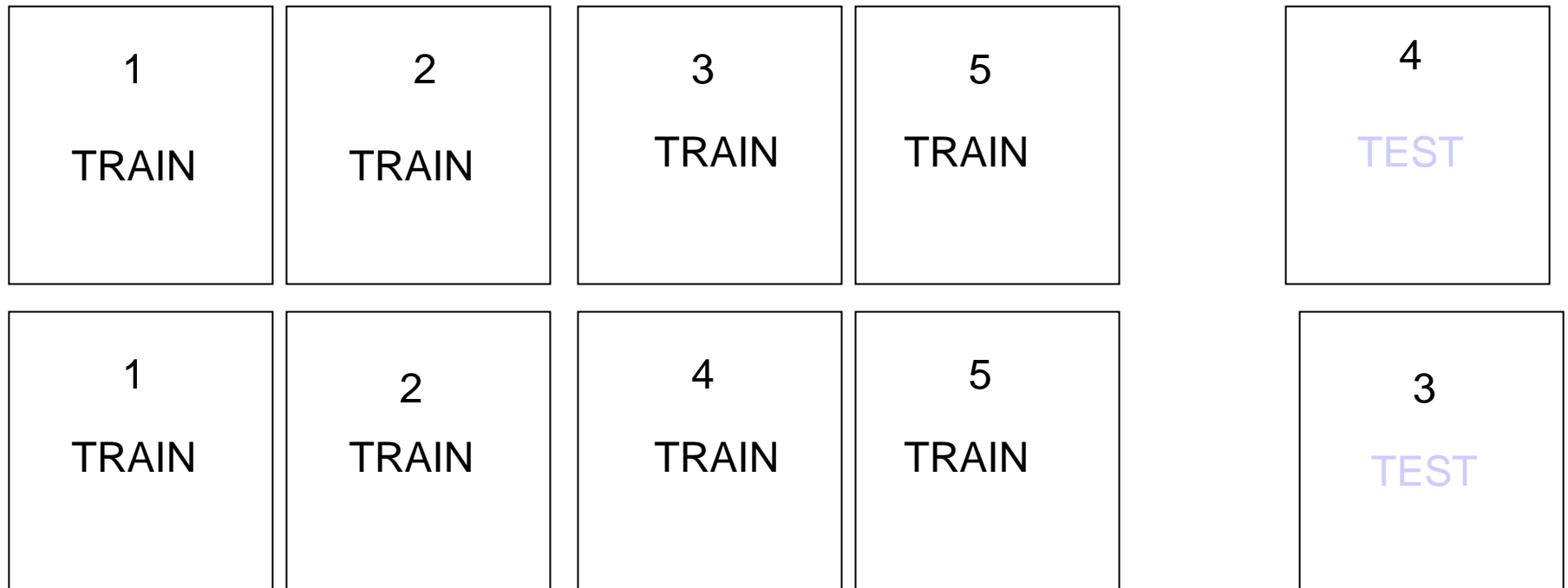


K-fold Cross Validation

For each part in turn, the classifier is built on the other K-1 parts then tested on the remaining part.



K-fold Cross Validation



etc....

Estimating Misclassification Error

- PAM estimates the predicted error rate based on misclassification error, which is calculated by averaging the errors from each of the cross validations.
- The model with lowest Misclassification Error is preferred.

Reducing the Feature Set

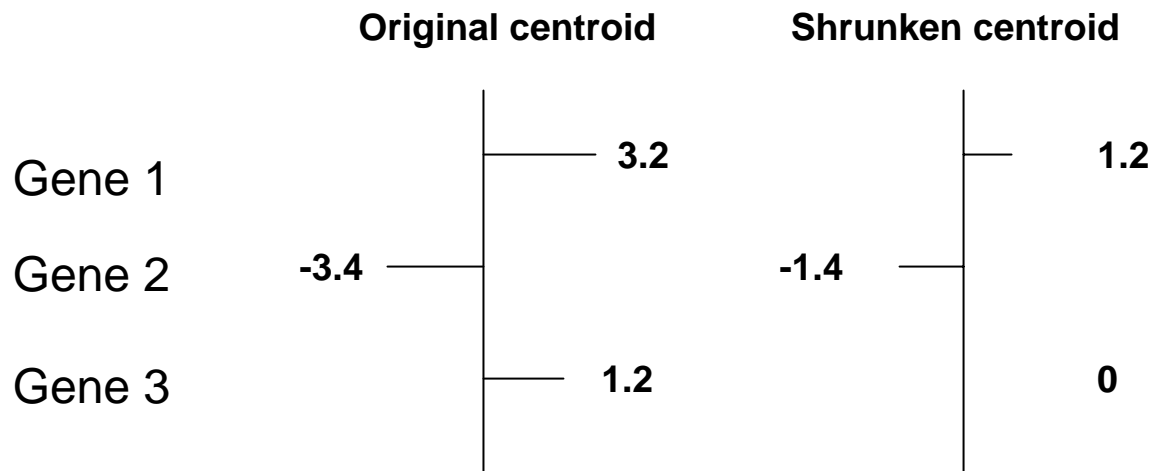
Nearest shrunken centroid classification makes one important modification to standard nearest centroid classification. It "shrinks" each of the class centroids toward the overall centroid for all classes by an amount we call the threshold . This shrinkage consists of moving the centroid towards zero by threshold, setting it equal to zero if it hits zero.

After shrinking the centroids, the new sample is classified by the usual nearest centroid rule, but using the shrunken class centroids.

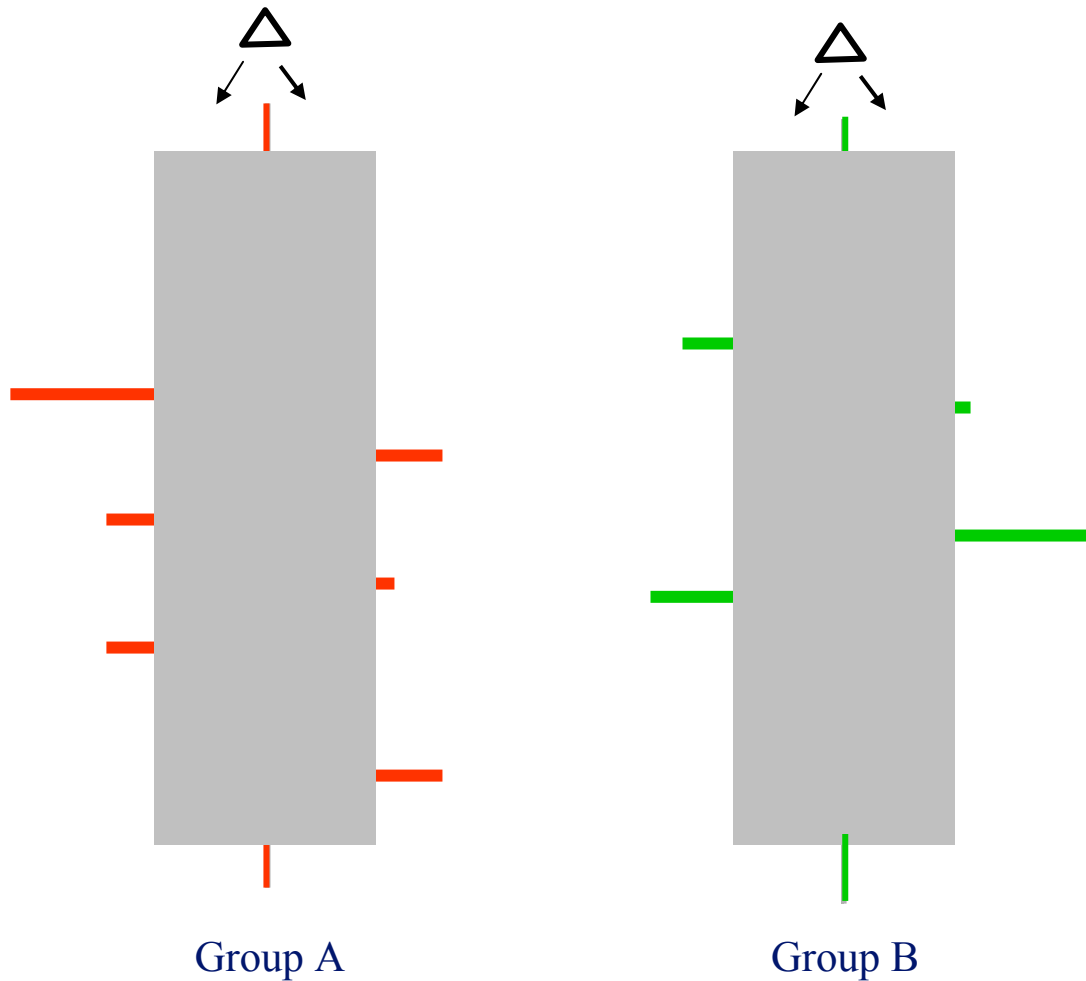
Shrinking the centroid

Threshold = 2.0:

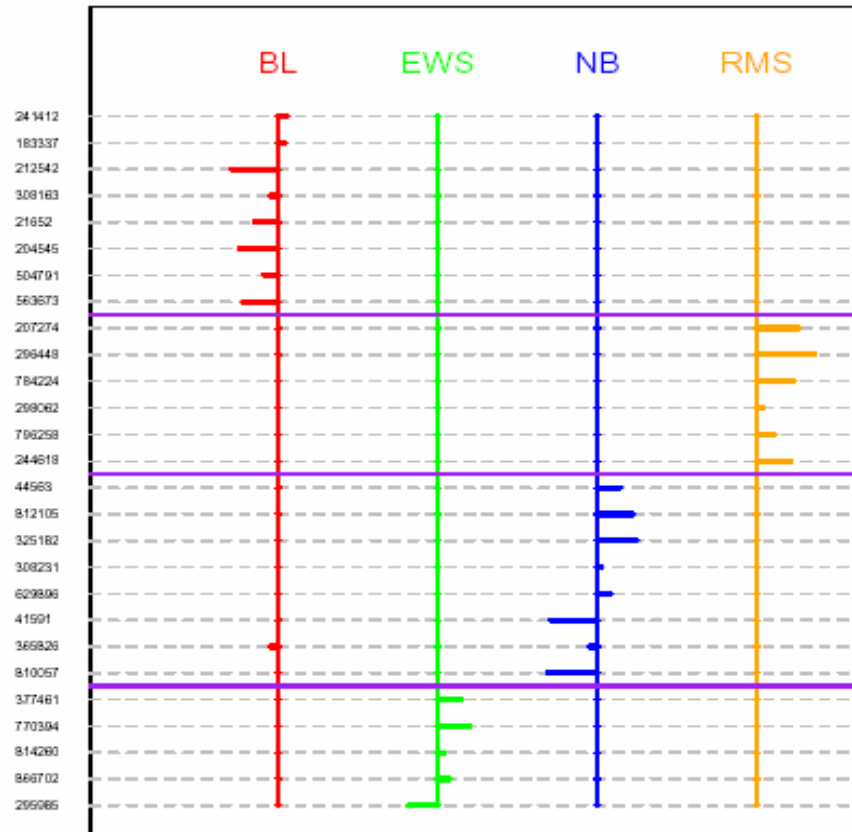
a centroid of 3.2 would be shrunk to 1.2;
a centroid of -3.4 would be shrunk to -1.4;
and a centroid of 1.2 would be shrunk to 0.



Reduce Gene Number

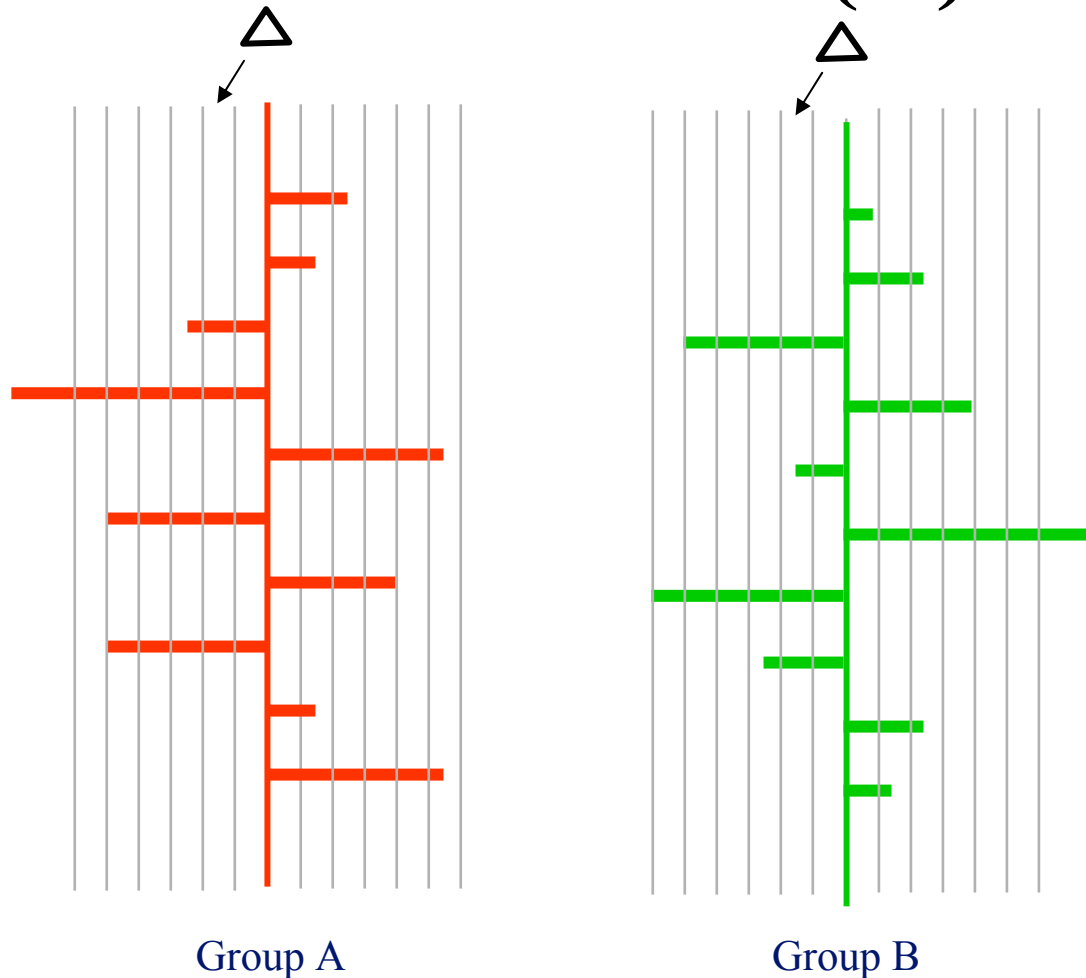


Prediction Model for SRBCT



- Compare model with new tumor tissues to make diagnosis

Multiple models with incremental threshold (Δ)



Sample

- 63 Arrays representing 4 groups
 - BL (Burkitt Lymphoma, $n_1=8$)
 - EWS (Ewing, $n_2=23$)
 - NB (neuroblastoma, $n_3=12$)
 - RMS (rhabdomyosarcoma, $n_4=20$)
- There are 2308 features (distinct gene probes)
- No missing values in array data sets
- Each group has an aggregate expression profile
- An unknown can be compared to each tumor class profile to predict which class it most likely belong

PAM Results

Clicking on a Delta value creates a new data Subset or enter

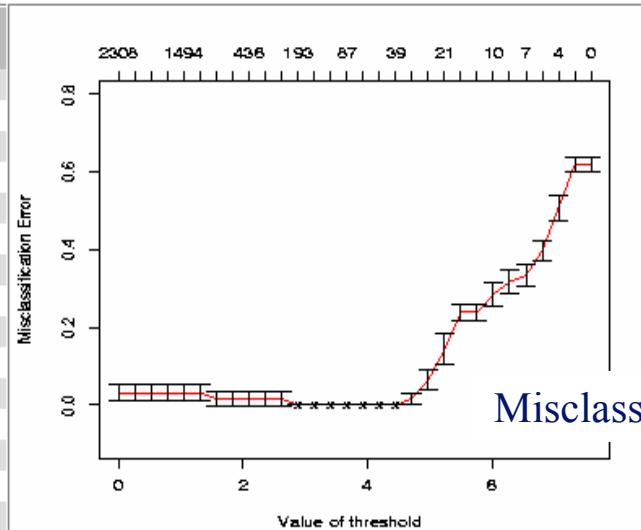
▼ a Delta value at the bottom and Click "Create Subset".

Shrinkage Delta	# of Genes	Misclass. Error
0.000	2308	0.032
0.262	2289	0.032
0.524	2145	0.032
0.786	1878	0.032
1.048	1494	0.032
1.309	1137	0.032
1.571	853	0.016
1.833	609	0.016
2.095	436	0.016
2.357	330	0.016
2.619	244	0.016
2.881 **	193	0.000
3.143 **	151	0.000
3.404 **	107	0.000
3.666 **	87	0.000
3.928 **	68	0.000
4.190 **	52	0.000
4.452 **	39	0.000
4.714	32	0.016
4.976	23	0.063
5.238	21	0.143
5.499	16	0.238
5.761	11	0.238
6.023	10	0.286
6.285	9	0.317
6.547	7	0.333
6.809	5	0.397
7.071	4	0.508

Link leads to the dataset
with PAM model →

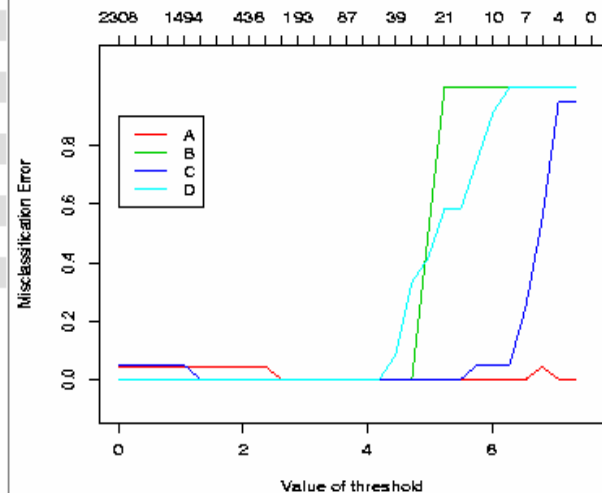
Create new model by fill
in a new Delta value →

Create Subset



Misclassification error

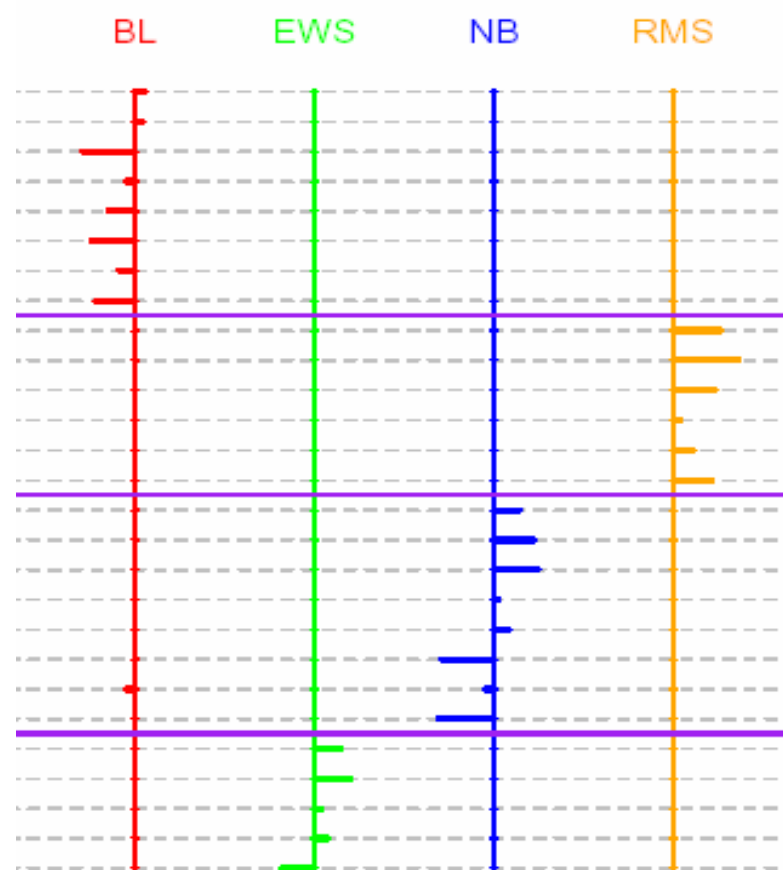
Above as [EPS](#), [PDF](#), [PNG](#)



mAdb PAM Model

↓ ↑	↓ ↑	↓ ↑	↓ ↑
A Score	B Score	C Score	D Score
0.6092	-0.0866	0.0000	0.0000
0.0000	0.0000	0.0000	0.5862
-0.0696	0.0000	0.0000	0.5764
-0.5421	0.0000	0.0000	0.0000
0.5338	0.0000	0.0000	0.0000
0.0000	-0.5321	0.0000	0.0000
0.0000	0.0000	0.0000	0.4936
0.0000	-0.4873	0.0000	0.0000
0.0000	0.0000	0.0000	0.4821
0.0000	-0.4661	0.0000	0.0000
0.4380	0.0000	0.0000	0.0000
-0.0110	0.0000	0.0000	0.4269
0.0000	-0.4153	0.0000	0.0000
0.4086	0.0000	0.0000	0.0000
0.0000	0.0000	-0.3828	0.0000
0.3346	0.0000	0.0000	0.0000

=



PAM summary

- It generates models (classifiers) from microarray data with phenotype information
- It does automatic gene selection for each models.
- Misclassification errors are calculated with the data for model selection.
- Require adequate numbers of samples in each group

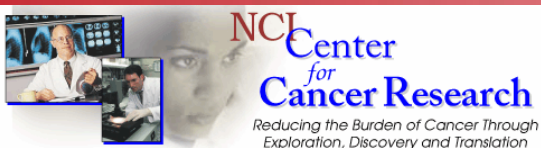
Hands-on Session 6

- Lab 11, Lab 12 (optional)
- Total time: 15 minutes

mAdb Development and Support Team:

- **John Powell, Chief, BIMAS, CIT**
- **Liming Yang, Ph.D**
- **Jim Tomlin**
- **Esther Asaki***
- **Yiwen He, Ph.D.***
- **Kathleen Meyer***
- **Tim Ruppert***

***SRA International contractor**



<http://madb.nci.nih.gov>
<http://madb.niaid.nih.gov>

For assistance, remember:

madb_support@bimas.cit.nih.gov

